

**UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES**

**MÉTHODES D'APPRENTISSAGE AUTOMATIQUE À FAIBLE  
COMPLEXITÉ POUR LA RECONNAISSANCE DE GESTES INTER-  
SESSIONS ET INTER-SUJETS À PARTIR DE HD-sEMG**

**THÈSE PRÉSENTÉE  
COMME EXIGENCE PARTIELLE DE LA  
DOCTORAT EN GÉNIE ÉLECTRIQUE**

**PAR  
MD RABIUL ISLAM**

**MARS 2024**

Université du Québec à Trois-Rivières

Service de la bibliothèque

Avertissement

L'auteur de ce mémoire, de cette thèse ou de cet essai a autorisé l'Université du Québec à Trois-Rivières à diffuser, à des fins non lucratives, une copie de son mémoire, de sa thèse ou de son essai.

Cette diffusion n'entraîne pas une renonciation de la part de l'auteur à ses droits de propriété intellectuelle, incluant le droit d'auteur, sur ce mémoire, cette thèse ou cet essai. Notamment, la reproduction ou la publication de la totalité ou d'une partie importante de ce mémoire, de cette thèse et de son essai requiert son autorisation.

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

LOW-COMPLEXITY MACHINE LEARNING METHODS FOR INTER-  
SESSION/SUBJECT GESTURE RECOGNITION FROM HD-sEMG

A THESIS PRESENTED  
IN PARTIAL FULFILLMENT OF  
THE REQUIREMENTS FOR THE DEGREE

DOCTOR OF PHILOSOPHY IN ELECTRICAL ENGINEERING

BY  
MD RABIUL ISLAM

MARCH 2024

UNIVERSITÉ DU QUÉBEC À TROIS-RIVIÈRES

DOCTORAT EN GÉNIE ÉLECTRIQUE (PH.D.)

**Direction de recherche :**

Daniel Massicotte, UQTR

Prénom et nom

directeur de recherche

Wei Ping-Zhu, Concordia University

Prénom et nom

codirecteur de recherche

**Jury d'évaluation**

Daniel Massicotte, UQTR

Prénom et nom

Directeur de recherche

Wei Ping-Zhu, Concordia University

Prénom et nom

Codirecteur de recherche

Usef Faghihi, UQTR

Prénom et nom

Membre interne

Messaoud Ahmed Ouameur, UQTR

Prénom et nom

Président du jury

Ahmed Lakhssassi, UQO

Prénom et nom

Membre externe

Thèse soutenue le 15 février 2024

# RÉSUMÉ

La reconnaissance des gestes à l'aide d'images à faible résolution de signaux électromyographiques de surface à haute densité (HD-sEMG) et à résolution instantanée ouvre de nouvelles perspectives pour le développement d'interfaces ordinateur-muscle plus fluides et naturelles. Les méthodes actuelles de pointe utilisent des réseaux de neurones convolutionnels (ConvNet) complexes, profonds et larges, ou un ensemble de ces réseaux complexes pour la reconnaissance des images HD-sEMG. Ce qui nécessite que l'architecture du réseau soit préentraînée sur un ensemble de données d'entraînement étiquetées à grande échelle. Par conséquent, cela rend l'application en temps réel coûteuse en ressources et en puissance de calcul.

Pour résoudre ce problème, les modèles légers S-ConvNet et All-ConvNet sont proposés, offrant ainsi un cadre simple, mais efficace pour apprendre les images HD-sEMG instantanées à partir de zéro pour la reconnaissance des gestes. Les résultats des expériences ont prouvé que les modèles proposés sont très efficaces pour apprendre des caractéristiques discriminatives pour la reconnaissance d'images HD-sEMG instantanées, en particulier dans des scénarios où les ressources et les données de qualité sont limitées. Les expériences menées sur quatre (4) ensembles de données HD-sEMG disponibles publiquement, sans utiliser des modèles préentraînés, ont montré que les modèles S-ConvNet et All-ConvNet proposés présentent une précision de reconnaissance des gestes intrasession très compétitive par rapport aux méthodes de pointes actuelles plus complexes en termes de paramètres d'apprentissage. Ainsi, les modèles S-ConvNet et All-ConvNet

proposés ont un grand potentiel pour apprendre des représentations discriminatives permettant de reconnaître les activités neuromusculaires sur des appareils aux ressources matériels limitées.

De plus, la variabilité des données entre les scénarios intersession et intersujet présente un défi important. Les approches existantes utilisent des réseaux de neurones convolutionnels complexes et profonds ou sur 2SRNN (2 étages de réseaux de neurones récurrents) basés sur des méthodes d'adaptation de domaine pour approximer le décalage de distribution causé par cette variabilité des données intersession et intersujet. Par conséquent, ces méthodes nécessitent également l'apprentissage de millions de paramètres d'entraînement et d'un ensemble de données de domaine source préentraîné et cible à la fois dans les étapes de préentraînement et d'adaptation. Par conséquent, le déploiement de ces méthodes gourmandes en ressources et coûteuses en puissance de calcul devient un défi pour les applications en temps réel.

Pour résoudre ce problème, l'adaptation de domaine (AD) avec S-ConvNet est proposée. Les méthodes d'AD proposées avec S-ConvNet apprennent une représentation transférable sur l'ensemble de données HD-sEMG de domaine source et les adaptent au domaine cible, même avec très peu de données disponibles, démontrant ainsi des capacités de généralisation améliorées en cas de décalage de distribution. Pour mieux résoudre ce problème, un modèle All-ConvNet+TL léger est proposé, qui exploite un ensemble de neurones convolutionnels légers et l'apprentissage par transfert (TL) pour améliorer les performances de reconnaissance de gestes intersession et intersujet. Le modèle All-ConvNet+TL se compose uniquement de couches convolutionnelles, offrant ainsi un cadre simple, mais efficace pour apprendre des représentations invariantes et discriminatives pour

résoudre les décalages de distribution dus à la variabilité des données intersession et intersujet. Des expériences menées sur quatre ensembles de données ont montré que la méthode d'AD proposée avec S-ConvNet, ainsi que le modèle All-ConvNet+TL, surpasse largement les approches existantes les plus complexes et établissent un nouveau record en matière de reconnaissance de gestes basée sur l'HD-sEMG dans des scénarios intersession et intersujet. Ces écarts de performance augmentent encore plus par rapport à l'état actuel de l'art lorsque de faibles quantités de données (par exemple, un seul essai) sont disponibles dans le domaine cible pour l'adaptation. Ces résultats expérimentaux exceptionnels fournissent des preuves que les modèles de pointe actuels peuvent être surentraînés inutilement pour les tâches de reconnaissance de gestes basés sur l'HD-sEMG en cas de variation entre les sessions et les sujets.

En outre, dans une autre étude, nous examinons la question de l'extraction d'ensembles de caractéristiques distinctives, et proposons ainsi d'utiliser l'histogramme de gradients orientés (HOG) comme caractéristiques uniques pour la reconnaissance robuste de l'activité neuromusculaire, en adoptant des SVM (séparateur à vaste marge) appariés comme schéma de classification. Les résultats expérimentaux ont montré que le HOG représente des caractéristiques uniques à l'intérieur de l'image HD-sEMG instantanée et que, en ajustant finement les hyperparamètres des SVM appariés, une précision de reconnaissance comparable à celle des méthodes de pointe plus complexes peut être obtenue.

## Abstract

Gesture recognition using low-resolution instantaneous high-density surface electromyography (HD-sEMG) images opens up new avenues for the development of more fluid and natural muscle-computer interfaces (MCI). However, the current state-of-the-art methods employed very complex deep and wide convolutional neural networks (ConvNet) or an ensemble of these complex networks for HD-sEMG image recognition, which requires the network architecture to be *pre-trained* on a very large-scale labeled training dataset. As a result, it makes high-end resource-bounded and computationally very expensive for deployment in real-time applications.

To overcome this problem, the S-ConvNet and lightweight All-ConvNet models are proposed, providing a simple yet efficient framework for learning instantaneous HD-sEMG images from scratch for gesture recognition. The experimental results proved that the proposed models are highly effective for learning discriminative features for instantaneous HD-sEMG image recognition, especially in the data and high-end resource-constrained scenarios. Experiments conducted on four (4) publicly available HD-sEMG datasets without using any pre-trained models, the proposed S-ConvNet and All-ConvNet demonstrate state-of-the-art or very competitive *intra-session* gesture recognition accuracy to the more complex current state-of-the-art, while significantly reducing the learning parameters. Hence, the proposed S-ConvNet and All-ConvNet have great potential for



learning discriminative representation for recognizing neuromuscular activities on resource-bounded devices.

Moreover, the data variability between inter-session and inter-subject scenarios presents a great challenge. The existing approaches employed very large and complex deep ConvNet or (2-Stage Recurrent Neural Networks) 2SRNN-based domain adaptation methods to approximate the distribution shift caused by these *inter-session* and *inter-subject* data variability. Hence, these methods also require learning over millions of training parameters and a large pre-trained and target domain dataset in both the pre-training and adaptation stages. Therefore, deploying these high-end, resource-intensive, and computationally expensive methods becomes challenging for real-time applications.

To address this problem, domain adaptation (DA) with S-ConvNet is proposed. The proposed DA methods with S-ConvNet learn transferable representation on the source domain HD-sEMG dataset and adapt them to the target domain, even with very limited data available, thus demonstrating enhanced generalization capabilities under distribution shift. To further address the problem, a lightweight All-ConvNet+TL model is proposed that leverages lightweight All-ConvNet and transfer learning (TL) for the enhancement of inter-session and inter-subject gesture recognition performance. The All-ConvNet+TL model consists solely of convolutional layers, a simple yet efficient framework for learning invariant and discriminative representations to address the distribution shifts caused by *inter-session* and *inter-subject* data variability. Experiments on four datasets demonstrate that the proposed DA method with S-ConvNet, as well as the All-ConvNet+TL model outperform the most complex existing approaches by a large margin and set a new state-of-the-art result on *inter-session* and *inter-subject* scenarios for sEMG-based gesture

recognition. These performance gaps increase even more against the current state-of-the-art when a tiny amount (e.g., a single trial) of data is available in the target domain for adaptation. These outstanding experimental results provide evidence that the current state-of-the-art models may be overparameterized for sEMG-based inter-session and inter-subject gesture recognition tasks.

Furthermore, in another study, we investigate the question of extracting distinctive feature sets, and thus propose to use *Histogram of Oriented Gradients* (HOG) as unique features for robust neuromuscular activity recognition, adopting pairwise SVMs (Support Vector Machine) as the classification scheme. The experimental results proved that the HOG represents unique features inside the instantaneous HD-sEMG image and by fine-tuning the hyper-parameter of the pairwise SVMs, the recognition accuracy comparable to the more complex state-of-the-art methods can be achieved.

## Acknowledgements

I would like to extend sincere appreciation to my advisor, Prof. Daniel Massicotte, for his patient supervision and encouragement throughout this research. I thank him for providing me with helpful guidance, comments, criticisms, facilities, and support for the research. I would like to convey my special thanks to my co-advisor Prof. Wei-Ping Zhu of Concordia University for his valuable comments, criticisms, and support for the research.

I would also like to thank Membres du jury d'évaluation, Prof. Messaoud Ahmed Ouameur, Prof. Ahmed Lakhssassi and Prof. Usef Faghihi for carefully reading and commenting on my thesis.

I appreciate the Dept. of Electrical and Computer Engineering of UQTR for the high-quality education and research facilities I received. The Laboratory of Signal and System Integration (LSSI) has always been a stimulating environment for conducting research. I thank all the academic, technical, and administrative staff of the department.

I thank my fellow researchers at LSSI for providing a friendly research environment and technical discussions.

I would like to express my sincere gratitude to my family members, especially my parents, wife Asma, Tasnim, Raed, Mousumy, grandmothers, uncles, aunties and cousins for their unending love, care, encouragement, and support.

Lastly, I would like to express my appreciation to the Regroupement Stratégique en Microsystèmes du Québec (ReSMiQ) for awarding several competitive doctoral fellowships, including the Anas Hamoui Best PhD Student fellowship in 2018.

# Table of Contents

<b>RÉSUMÉ.....</b>	<b>iii</b>
<b>Abstract.....</b>	<b>vi</b>
<b>Acknowledgements.....</b>	<b>ix</b>
<b>Table of Contents .....</b>	<b>x</b>
<b>List of Tables .....</b>	<b>xv</b>
<b>List of Figures.....</b>	<b>xvii</b>
<b>List of Abbreviations .....</b>	<b>xx</b>
<b>Chapitre 1 - Introduction .....</b>	<b>1</b>
1.1 Background.....	1
1.2 Research Problems .....	4
1.3 Objectives.....	5
1.4 Research Methodology.....	6
1.5 Contribution of the study.....	11
1.6 Organization of the thesis.....	13
<b>Chapitre 2 - HD-sEMG-Based Gesture Recognition – State-of-the-Art.....</b>	<b>14</b>
2.1 Background.....	15
2.2 Feature Extraction and Classical Machine Learning (ML) Methods .....	18

2.3	Deep Learning Methods for sEMG-based Gesture Recognition.....	24
2.4	Inter-Session/Inter-Subject Scenarios .....	29
2.5	CapgMyo Dataset .....	32
2.6	Conclusion.....	35
<b>Chapitre 3 - Domain Adaptation with Low-Latency Shallow Convolutional</b>		
<b>Neural Networks for Improved Inter-Session/Inter-Subject Gesture</b>		
<b>Recognition .....</b>		
		<b>36</b>
3.1	Introduction .....	37
3.2	The Proposed Framework.....	42
3.3	Model Description– The Shallow Convolutional Neural Network (S- ConvNet) .....	43
	3.3.1 S-ConvNet Architecture and Training.....	46
	3.3.2 Normalization .....	50
3.4	The Performance Evaluation of The Proposed S-ConvNet Network Models .....	51
	3.4.1 Data Selection for Training, Validation and Testing.....	52
	3.4.2 Experimental Results on Intra-Session Scenarios .....	54
	3.4.3 Discussion on sEMG-based Gesture Recognition in Intra-Session Scenarios .....	66
3.5	Domain Adaptation with Low-Latency Shallow Convolutional Neural Network [100] .....	68

3.5.1 Preliminaries.....	70
3.5.2 Baseline DA Framework and Its Limitations.....	70
3.5.3 Proposed Domain Adaptation (DA) Framework.....	72
3.6 Experiments in Inter-Session and Inter-Subject Scenarios .....	74
3.6.1 Experimental set up.....	74
3.6.2 sEMG-Based Gesture Recognition in Inter-Session Scenarios.....	75
3.6.3 sEMG-Based Gesture Recognition in Inter-Subject Scenarios.....	77
3.6.4 Discussion on sEMG-Based Gesture Recognition in Inter- Session/Inter-Subject Scenarios.....	80
3.7 Conclusion.....	81
<b>Chapitre 4 - Surface EMG-Based Inter-Session/Inter-Subject Gesture Recognition by Leveraging Lightweight All-ConvNet and Transfer Learning.....</b>	<b>83</b>
4.1 Introduction .....	84
4.2 The Proposed Transfer Learning Framework.....	89
4.3 Model Description – The All-Convolutional Neural Network (All- ConvNet) .....	90
4.4 Transfer Learning by Leveraging Lightweight All-ConvNet (All-ConvNet+TL) .....	96
4.5 Experimental Setup .....	99

4.6	Experimental Results.....	101
4.6.1	Intra-Session Performance Evaluation .....	101
4.6.2	Inter-Session Performance Evaluation .....	108
4.6.3	Inter-Subject Performance Evaluation .....	111
4.6.4	Weight (or Feature) Transfusion Experiments .....	114
4.6.5	Lightweight All-ConvNet Network Trimming .....	115
4.7	Discussion.....	117
4.8	Conclusion.....	119
 <b>Chapitre 5 - HOG and Pairwise SVMs for Neuromuscular Activity</b>		
	<b>Recognition Using Instantaneous HD-sEMG Images.....</b>	<b>121</b>
5.1	Introduction .....	122
5.2	The Proposed Neuromuscular Feature Extraction and Classification Algorithm .....	123
5.2.1	Histogram of Oriented Gradients (HOG) Feature Extraction .....	124
5.2.2	Pairwise SVM Classifier .....	129
5.3	Experiments.....	129
5.4	Conclusions .....	132
 <b>Chapitre 6 - Conclusion .....</b>		
	<b>.....</b>	<b>134</b>
6.1	Summary.....	134
6.2	Future work and directions.....	138

**References ..... 141**

**Appendix A – Published Articles ..... 149**



## List of Tables

Table 2.1 Gestures in DB-a and DB-b (8 isotonic and isometric hand configurations). Adapted from [26].	34
Table 2.2 Gestures in DB-c 8 isotonic and isometric hand configurations). Adapted from [26].	34
Table 3.1 The three S-Convnet networks Models for sEMG-based gesture recognition using instantaneous sEMG images.	46
Table 3.2 Gesture recognition accuracy (%) using instantaneous HD-sEMG Images for different activation functions and spatial pooling.	55
Table 3.3 The average recognition accuracies (%) of 8 hand gestures for CapgMyo DB-a and DB-b for 18 and 10 different subjects respectively and 12 gestures for 10 different subjects in DB-c. The numbers are the majority voted results using 160 ms window (i.e., 160 frames). per-frame accuracies are shown in parenthesis.	59
Table 3.4 Inter-session gesture recognition accuracies on CapgMyo DB-b. The average recognition accuracies (%) of 8 hand gestures for 10 different subjects respectively. The numbers are the majority voted results using 150 ms window (i.e., 150 frames).	77
Table 3.5 Inter-subject gesture recognition accuracies. The average recognition accuracies (%) of 8 hand gestures for CapgMyo DB-b and 12 hand gestures for CapgMyo DB-c for 10 different subjects respectively. The numbers are the majority voted results using 150 ms window (i.e., 150 frames).	79
Table 3.6 Inter-session and Inter-subject improvement (%) results obtained by the proposed DA with S-ConvNet.	80
Table 4.1 The All-Convnet Network Model for Neuromuscular Activity Recognition.	94
Table 4.2 The average recognition accuracies (%) of 8 hand gestures for CapgMyo DB-a and DB-b for 18 and 10 different subjects respectively and 12 gestures for 10 different subjects in DB-c. The	

numbers are the majority voted results using 160 ms window (i.e., 160 frames). Per-frame accuracies are shown in parenthesis.....	103
Table 4.3 Inter-session gesture recognition accuracies on CapgMyo DB-b. The average recognition accuracies (%) of 8 hand gestures for 10 different subjects respectively. The numbers are the majority voted results using 150 ms window (i.e., 150 frames). .....	110
Table 4.4 Inter-subject gesture recognition accuracies. The average recognition accuracies (%) of 8 hand gestures for CapgMyo DB-b and 12 hand gestures for CapgMyo DB-c for 10 different subjects respectively. The numbers are the majority voted results using 150 ms window (i.e., 150 frames).....	112
Table 4.5 Inter-session and Inter-subject improvement (%) results obtained by the proposed lightweight All-ConvNet+TL leveraging transfer learning. ....	113
Table 4.6 Inter-session and Inter-subject gesture recognition accuracies (%) under full feature extraction setting.....	114
Table 4.7 Learning (or convergence) speed using various training epochs. Table shows inter-session gesture recognition accuracies (%) on test set. The numbers are the majority voted results using 150 ms window (i.e., 150 frames). Per-frame accuracies are shown in parenthesis.....	116
Table 4.8 Learning (or convergence) speed using various training epochs. Table shows inter-session gesture recognition accuracies (%) on test set. The numbers are the majority voted results using 150 ms window (i.e., 150 frames). Per-frame accuracies are shown in parenthesis.....	116
Table 5.1 Precision and Recall of every gesture classes. ....	132

## List of Figures

Fig. 2.1	Generation and decomposition mechanism of sEMG signals. Adapted from [96].	16
Fig. 2.2	Schematic illustration of sEMG-based gesture recognition by windowing sEMG signals. Adapted from [21]).	17
Fig. 2.3	Measurements of all pairs of distances at a) 1-point and b) 2-points apart along the $x$ -directions for an HD-sEMG grid. Adapted from [18].	23
Fig. 2.4	A schematic illustration of the ConvNet architecture employed by Geng et al. [21]. Adapted from [26].	25
Fig. 2.5	A schematic illustration of multi-stream decomposition and fusion network. Adapted from [23].	26
Fig. 2.6	A schematic elaboration of CNN-RNN network with an attention module for sEMG-based gesture recognition. Adapted from [24].	27
Fig. 2.7	A schematic diagram of the 3D convolutional neural network architecture. Adapted from [36].	28
Fig. 3.1	Schematic diagram of the proposed framework of muscular activity recognition by instantaneous sEMG images.	43
Fig. 3.2	Total number HD-sEMG images seen during training, for pre-training + fine-tuning vs. random-initialization.	44
Fig. 3.3	A schematic illustration of convolutions and pooling operation a) Convolution maps and b) Convolutions maps after spatial pooling.	48
Fig. 3.4	The average per-frame gesture recognition accuracy of 8 hand gestures for 18 different subjects in CapgMyo DB-a with the proposed S-ConvNet models and the current state-of-the-art GengNet [21], [26] model.	60
Fig. 3.5	The per-frame gesture recognition accuracy with the proposed S-ConvNet A (a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, (b)-(c) The gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo DB-b (Session 1) and DB-b (Session 2) respectively (d) the gesture	

recognition accuracy of 12 hand gestures for 10 different subjects on CapgMyo DB-c. .... 62

Fig. 3.6 Surface EMG gesture recognition accuracy with different voting windows using the proposed S-ConvNet models and compared with the state-of-the-art methods: (a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, (b)-(c) the gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo DB-b (Session 1) and DB-b (Session 2) respectively, and (d) the recognition accuracy of 12 hand gestures for 10 different subjects on DB-c. .... 65

Fig. 3.7 A schematic illustration of the proposed DA using shallow convolutional neural network (S-ConvNet). (a) Pre-trained model (b) Proposed DA Framework. sEMG images and labels used for DA are shown. .... 73

Fig. 4.1 A general conceptual diagram of the transfer learning method (a) Pre-trained model (b) Fine-tuned model and (c) Feature extraction process. sEMG images and labels used for adaptation are shown. .... 91

Fig. 4.2 HD-sEMGs derived from the same muscular activity class which demonstrates that the distributions are independent to the class labels. .... 93

Fig. 4.3 A schematic illustration of feature maps obtained by All-ConvNet before and after dimensionality reduction. (a) Feature maps and b) Feature maps after dimensionality reduction. .... 95

Fig. 4.4 The per-frame gesture recognition accuracy with our proposed lightweight All-ConvNet (a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, (b)-(c) The gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo DB-b (Session 1) and DB-b (Session 2) respectively (d) the gesture recognition accuracy of 12 hand gestures for 10 different subjects on CapgMyo DB-c. .... 105

Fig. 4.5 Surface EMG gesture recognition accuracy with different voting windows using the proposed lightweight All-ConvNet and compared with the state-of-the-art methods: a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, and the gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo for b) DB-b Session 1 and c) DB-b Session 2, and d) the recognition accuracy of 12 hand gestures for 10 different subjects on DB-c. .... 107

Fig. 5. 1 Schematic illustration of the proposed muscular activity recognition by instantaneous sEMG images. .... 124

- Fig. 5.2 HOG extraction process (a) An instantaneous sEMG image is partitioned by non-overlapping cells and overlapping blocks (each block has  $(2 \times 2)$  four cells). (b) Gradients information are overlaid over an instantaneous sEMG image (c) HOG in each block. The horizontal axis represents angle information and the vertical axis bears weighted histogram..... 127
- Fig. 5.3 Confusion Matrix of the Proposed Neuromuscular Activity Recognition Method..... 132

## List of Abbreviations

EMG	Electromyography
sEMG	Surface Electromyography
MCI	Muscle-Computer Interface
HD-sEMG	High-Density Surface Electromyography
ML	Machine Learning
DL	Deep Learning
DNN	Deep Neural Networks
LDA	Linear Discriminant Analysis
SVM	Support Vector Machines
HMM	Hidden Markov Model
GMM	Gaussian Mixture Model
CNN	Convolutional Neural Network
ConvNet	Convolutional Neural Network
RNN	Recurrent Neural Network
2SRNN	2-stage Recurrent Neural Network
DA	Domain Adaptation
S-ConvNet	Shallow-Convolutional Neural Network
All-ConvNet	All Convolutional Neural Network
TL	Transfer Learning
HOG	Histogram of Oriented Gradients
HCI	Human Computer Interface
MU	Motor Unit
MUAP	Motor Unit Action Potential
MLP	Multi-Layer Perceptron
KNN	K-Nearest Neighbor
RMS	Root Mean Square
iEMG	Integrated EMG

ZC	Zero-Crossing
WL	Waveform Length
VAR	Variance
MAV	Mean absolute value
PS	Power Spectrum
MNF	Average Frequency
MDF	Intermediate Frequency
FR	Frequency Ratio
CC	Cepstrum Coefficients
AR	Autoregressive Coefficient
FFT	Fast Fourier transform
MF	Median Frequency
MPF	Mean Power Frequency
DWTC	Discrete Wavelet Transform Coefficients
DWPTC	Discrete Wavelet Packet Transform Coefficients
CWTC	Continuous Wavelet Transform Coefficients
TSD	Temporal-Spatial Descriptor
BN	Batch Normalization
AdaBN	Adaptive Batch Normalization
SGD	Stochastic Gradient Descent
Adam	Adaptive Moment Estimation
LOTOV	Leave-One-Trial-Out Cross-Validation
LOSOV	Leave-One-Subject-Out Cross-Validation
CPU	Central Processing Unit
GPU	Graphics Processing Unit
TPR	True Positive Rate or Recall
PPV	Positive Predictive Value or Precision

# Chapitre 1 - Introduction

## 1.1 Background

Gesture or neuromuscular activity recognition based on surface electromyography (sEMG) signals has been a core technology for developing next-generation muscle-computer interfaces (MCIs). The major application domains of sEMG-based MCIs are non-intrusive control of active prosthesis [1], wheelchairs [2], exoskeletons [3] or neurorehabilitation [4], neuromuscular diagnosis [5] and providing interaction methods for video games [6], [7]. The conventional approaches for gesture recognition using sparse multi-channel sEMG sensors and classical machine learning methods – such as linear discriminant analysis (LDA) [8], support vector machines (SVM) [9], hidden Markov model (HMM) [10] – on windowed descriptive and discriminative time-domain, frequency-domain and/or time-frequency-domain sEMG feature space [11], [12-16]. However, these sparse multi-channel sEMG-based methods are not suitable for real-world applications due to their lack of robustness to electrode shift and positioning [17], [18]. In addition, malfunction to any of these sparse-channel electrodes leads to retraining the entire MCI system. Deep learning-based methods have recently been exploited for gesture recognition using sparse multi-channel sEMG [19-20], [31-32], [61] but their performance is still far from optimum [64].

To address this problem, designing and developing more flexible, convenient, and comfortable high-density sEMG (HD-sEMG) based myoelectric sensors and efficient



pattern recognition algorithms have been major research directions in recent years [17-18], [21-30], [36]. The HD-sEMG records myoelectric signals using two-dimensional (2D) electrode arrays that characterize the spatial distribution of myoelectric activity over the muscles that reside within the electrode pick-up area [21]. The collected HD-sEMG data are spatially correlated which enabled both temporal and spatial changes and robust against malfunction of the channels with respect to the previous counterparts [18]. However, the existing HD-sEMG-based gesture recognition methods [17-18], [28], [30] still rely on the windowed sEMG (e.g., range between 100 ms and 300 ms [33], [34]), which demands finding an optimal window length. The determination of an optimal window length represents a strong trade-off between classification accuracy and controller delay, both of which increase with an increase in window size.

To further address this problem, distinctive patterns within instantaneous sEMG images were first discovered by Geng *et al.* [21] and Du *et.al.* [26] to develop more fluid and natural muscle-computer interfaces (MCIs). The instantaneous values of HD-sEMG signals at each sampling instant were arranged in a 2D grid in accordance with the electrode positioning. Subsequently, this 2D grid was transformed into a grayscale sEMG image. Therefore, an instantaneous sEMG image represents a relative global measure of the physiological processes underlying neuromuscular activities at a given time. Consequently, gesture recognition is performed solely with the sEMG images spatially composed from HD-sEMG signals recorded at a specific instant. Hence, the observational latency was reduced to only 1 ms, which would significantly decrease controller delay for the above-mentioned MCI applications.

Geng *et al.* [21] employed a deep convolutional neural network (CNN or ConvNet) to recognize hand gestures from the sEMG images and showed high recognition accuracy on publicly available benchmark HD-sEMG datasets [15], [17], [26]. Du *et al.* [26] employed the same ConvNet as proposed in [21]; however, they applied adaptive batch normalization to this ConvNet to enhance the scalability of the classifier for sEMG-based gesture recognition. Motivated by these prior works, further studies have been conducted on this promising new research direction over the years [22-25], [27], [29], [36].

However, the state-of-the-art methods [21], [23], [24], [26], [61] for sEMG-based gesture recognition either employed very complex deep and wide CNN or an ensemble of these complex networks for improved gesture recognition performance. Despite the significant performance boost achieved by these state-of-the-art models [21], [23], [24], [26], [61] the heavy computational and intensive memory cost hinders deploying them on resource-constrained embedded and mobile devices for real-time applications. Therefore, the demand for designing low-cost, low-latency and lightweight networks is highly increasing for low-end resource-limited embedded and mobile devices.

Moreover, the problem of sEMG-based gesture recognition becomes significantly more challenging in operational conditions or in *inter-session/inter-subject* scenarios, where a trained model is deployed to recognize muscular activities in a new recording session involving the same subject or when encountering completely new or unseen subjects (i.e., inter-subject) [1], [26], [63].

Furthermore, the existing sEMG-based gesture recognition methods are typically evaluated using data acquired from able-bodied subjects. However, it is important to note that sEMG signals are highly specific to each individual, and in real-time sEMG-based MCI

applications such as assistive technology and physical rehabilitation [1-5], the target users are often elderly individuals, amputees, and patients. These differences between the source domain task (able-bodied subjects) and the target domain task (individuals with motor control impairments) pose a significant challenge [26], [63]. The main issue arises from the fact that sEMG-based gesture recognition in the target domain needs to be conducted with limited data availability due to the difficulty of acquiring data from amputees, elderly individuals, patients, and similar groups.

## 1.2 Research Problems

The current state-of-the-art methods [21], [23], [24], [26], [61] for sEMG-based gesture or neuromuscular activity recognition either employed very complex deep and wide CNN or an ensemble of these complex networks for improved *intra-session* (i.e., where train and test sEMG signals are recorded at the same session) gesture recognition performance. Therefore, these methods require learning over millions of training parameters and large-scale labeled HD-sEMG training datasets for pre-training, making them computationally expensive and resource-intensive.

The sEMG-based gesture recognition problem becomes more challenging in operational conditions or an inter-session scenario, wherein a trained model is deployed to recognize muscular activities in a new recording session for the same subject. The distributions of the sEMG signals captured in a new recording session deviate from those obtained during the training session due to electrode shifts, variations in arm posture, and time-dependent changes like fatigue and electrode-skin contact impedance [1], [17], [26], [63]. Inter-session is referred to as inter-subject in cases where training and test data are obtained from different subjects [26]. However, in this thesis, inter-subject is considered only in cases

where training and test data are obtained from different subjects. In the inter-subject scenario, the data variability comes from the variation in muscle physiology between different subjects. Moreover, it is always challenging to force the users to maintain a certain level of muscular contraction force in real-time applications. Therefore, the developed methods must also cope with the distribution shift that occurred by this voluntary muscular contraction force level in addition to the distribution shift caused by inter-session and inter-subject data variability.

To approximate the distribution shift caused by these inter-session and inter-subject data variability, the current state-of-the-art methods [26], [57] employed very large and complex deep ConvNet or 2-Stage Recurrent Neural Networks (2SRNN)-based domain adaptation (DA) methods. Hence, these methods also require learning over millions of training parameters and a large pre-trained and target domain dataset in both the pre-training and adaptation stages. Therefore, deploying these high-end, resource-constrained, and computationally expensive methods becomes challenging for real-time applications. Also, the large computationally expensive models might significantly impede mobile and on-device applications, where power consumption, data memory, and computational speed are constraints.

Therefore, designing and developing efficient and resource-constrained robust domain-invariant feature representations and classification techniques is highly demanded to accurately decode and discriminate movements for sEMG-based gesture recognition.

### **1.3 Objectives**

The main objective of this project is to develop efficient and resource-constrained robust domain-invariant non-invasively feature representations and classification techniques for

improved sEMG-based gesture recognition using instantaneous values of HD-sEMG signals. To obtain this objective, the following sub-objectives are pursued:

- (i) Propose S-ConvNet: A shallow convolutional neural network architecture for sEMG-based gesture recognition using instantaneous HD-sEMG images.
- (ii) Propose a domain adaptation method with low-latency shallow CNN to approximate the domain shift for enhancement of sEMG-based gesture recognition accuracy.
- (iii) Propose All-ConvNet: A lightweight All-CNN for sEMG-based gesture recognition using instantaneous HD-sEMG images.
- (iv) Introduce the All-ConvNet+TL model, a novel framework which leverages the lightweight All-ConvNet and transfer learning to address the distribution shift in inter-session and inter-subject sEMG-based gesture recognition.
- (v) Introduce a network pruning/trimming method based on the findings of the proposed weight (or feature) transfusion experiments to further optimize the proposed lightweight All-ConvNet+TL model. This involves selectively pruning the network's weights, leading to developing of a more efficient Lightweight All-ConvNet-Slim model.
- (vi) Study the question of extracting distinctive feature sets and propose to use Histogram of Oriented Gradients (HOG) as unique features for HD-sEMG image recognition, adopting pairwise SVMs as the classification scheme.

#### **1.4 Research Methodology**

This research is conducted in a systematic order. The research is carried out starting from literature review to identify the major problems in the current state-of-the-art methods and

their potential solutions for sEMG-based gesture recognition using instantaneous HD-sEMG images. The current state-of-the-art methods [21], [23], [24], [26], [36], [61] either employed very complex deep and wide CNN or an ensemble of these complex networks for improved gesture recognition performance. For example, Geng *et al.* [21] and Du *et al.* [26] exploited a DeepFace [35] like very large and deep CNN (dubbed as GengNet), which requires learning  $>5.63M$  (million) training parameters only during fine-tuning and pre-trained on a very large-scale labeled sEMG training datasets. The complexity of this model grows linearly as the input size is increased due to the use of an unshared weight strategy [45], [27]. Wei *et al.* [23] employed an ensemble of eight (8) single-stream GengNet within their gesture recognition framework based on instantaneous HD-sEMG images. An ensemble of multi-stream GengNet and long-short term memory networks (LSTM) augmented with an attention module is proposed in [24]. Chen *et al.* [36] employed 3D CNN for learning spatial and temporal representation of sEMG images. However, the employed 3D CNN requires learning over at least  $> 30 M$  parameters, which is not feasible for real-time MCI applications based on sEMG signals. Hence, these methods are not feasible for deploying on resource-constrained mobile and embedded devices for real-time applications.

To address these challenges, this thesis introduces low-latency and parameter efficient shallow convolutional neural networks (S-ConvNet) model architectures, specifically targeting sEMG-based gesture recognition on resource-constrained low-end devices. S-ConvNet is designed to learn sEMG image representation from scratch through random initialization. S-ConvNet consists of a network with simple convolution layers with the shared kernel, a fully connected layer with a small number of neurons, and an occasional

dimensionality reduction performed by stridden CNN, demonstrating state-of-the-art recognition accuracy on publicly available benchmark HD-sEMG datasets, while needing to be learnt  $\approx 1/4$ th learning parameters using a  $\approx 12 \times$  smaller dataset outperforming the more complex and high-end resource-bounded state-of-the-art methods.

In addition, striving to find a simpler and more efficient lightweight network, a new architecture called All-ConvNet is introduced that consists solely of convolutional layers and is designed to be more efficient and less computationally intensive than the existing state-of-the-art models for sEMG-based gesture recognition. Comparing the performance of All-ConvNet to other state-of-the-art models shows that it achieves competitive or state-of-the-art performance on current benchmark HD-sEMG datasets, while being significantly lighter, more efficient, and faster to train and evaluate. *The design of All-ConvNet was motivated by the finding of fact that when the units in the uppermost convolutional layer adequately encompass a significant region of the sEMG image, it becomes feasible to accurately recognize the content of the image, specifically the targeted gesture class.* This results in the generation of predictions for sEMG image classes at various positions, which can then be easily averaged across the entire image. Hence, the proposed All-ConvNet becomes robust to translations and geometric distortions, which can be very effective in addressing the challenging electrode shift and positioning as well as electrodes malfunctioning problem in sEMG-based gesture recognition.

Moreover, the existing methods (e.g., [21], [26]) reported gesture recognition rate as low as 20% using the conventional classifiers such as support vector machines (SVM). However, in another study, this thesis argued that the conventional classifiers such as SVM can surpass ConvNet at producing optimal classification if well-behaved feature vectors are

provided. To present an alternative solution to the resource-constrained ConvNet, this thesis also delves into the question of extracting distinctive feature sets. This thesis propose to use Histogram of Oriented Gradients (HOG) as distinctive features for robust gesture or neuromuscular activity recognition, adopting pairwise SVMs as the classification scheme.

Furthermore, the data variability between inter-session and inter-subject scenarios presents a great challenge. The current-state-of-the-art methods [26], [57] employed very large and complex deep ConvNet or 2SRNN-based domain adaptation methods to approximate the distribution shift caused by these inter-session and inter-subject data variability. Hence, these methods require learning over millions of training parameters and a large pre-trained and target domain dataset in both the pre-training and adaptation stages. Therefore, deploying these high-end, resource-constrained, and computationally expensive methods becomes challenging for real-time applications.

To address this distribution shift problem, domain adaptation (DA) with shallow convolutional neural network (S-ConvNet) is proposed. DA leverages S-ConvNet to learn transferable representations in an efficient manner from a source domain task (or dataset) to target domain task (or dataset).

To further simplify the distribution shift problem caused by inter-session and inter-subject data variability, a lightweight All-ConvNet+TL model is proposed by leveraging lightweight All-ConvNet and transfer learning (TL) for the enhancement of *inter-session* and *inter-subject* gesture recognition performance. The All-ConvNet+TL model consists solely of convolutional layers, a simple yet efficient framework for learning domain-invariant and discriminative representations. The proposed DA method with S-ConvNet as well as All-ConvNet+TL outperformed the current state-of-the-art DA methods by a large



margin both when the data from single trials or multiple trials are available for domain adaptation.

To achieve optimal performance by All-ConvNet+TL, a weight (or feature) transfusion experiment is conducted involving the partial reuse of pre-trained weights, the specific regions where valuable feature reuse manifests is uncovered and delve into the exploration of hybrid approaches for transfer learning. These approaches involve using a subset of pre-trained weights and redesigning other parts of the network to make them more lightweight.

Experiments and analysis were also conducted to justify the potentiality of these proposed methods. A series of experiments were carried out on four (4) publicly available benchmark HD-sEMG datasets [26] for sEMG-based gesture recognition in intra-session, inter-session and inter-subject scenarios.

Finally, the proposed research methodology is summarized as follows: 1) Conduct an extensive literature review for sEMG-based gesture recognition in intra-session, inter-session, and inter-subject scenarios. The research problems and major drawbacks in the current state-of-the-art solutions are identified, 2) Explore publicly available HD-sEMG datasets, and prepare the data scenarios, 3) Design and develop low-latency S-ConvNet model architectures targeting sEMG-based gesture recognition on low-end devices. Introduce a domain adaptation method with low-latency S-ConvNet to address the domain shift problem of sEMG-based gesture recognition in inter-session and inter-subject scenarios, 4) Design and develop a low-latency, memory/parameter efficient lightweight All-ConvNet. Explore an efficient transfer learning framework by leveraging lightweight All-ConvNet to address distribution shifts caused by inter-session and inter-subject data variability, 5) Introduce a novel HoG-based feature extraction method for sEMG-based

gesture recognition with classical machine learning methods, 7) Finally, compare the results with the current state-of-the-art methods using the same datasets and scenarios.

### 1.5 Contribution of the study

Based on the proposed methodology and research approach described, the main contributions of this thesis are as follows:

1. Proposed low-latency S-ConvNet [25], [63], a shallow convolutional neural network architecture that achieved state-of-the-art performance on four (4) publicly available benchmark HD-sEMG datasets for *intra-session* gesture recognition tasks.
2. Proposed a domain adaptation method with low-latency shallow convolutional neural network to approximate the domain shift and achieved state-of-the-art results for *inter-session* and *inter-subject* gesture recognition tasks [100].
3. Proposed a low-latency and parameter/memory efficient All-ConvNet [27], A lightweight All-CNN that demonstrates highly competitive or even outperforms the most complex state-of-the-art methods, for sEMG-based *intra-session* gesture recognition tasks on a current benchmark HD-sEMG dataset.
4. Introduced All-ConvNet+TL model [67], which leverages the lightweight All-ConvNet and transfer learning to address the distribution shift in *inter-session* and *inter-subject* sEMG-based gesture recognition and evaluate it against the more complex state-of-the-art models. The All-ConvNet+TL model outperforms the state-of-the-art models by a large margin, both when the data from a single trial or multiple trials are available for fine-tuning/adaptation.

5. A weight (or feature) transfusion experiment [67] is introduced, where we partially reuse pre-trained weights. We uncover specific regions where valuable feature reuse manifests and delve into the exploration of hybrid approaches for transfer learning. To the author's knowledge, this is the first study to conduct weight transfusion experiments for sEMG-based gesture recognition.
6. Designing an efficient lightweight ConvNet architecture is crucial for optimizing computational and memory costs. Building upon the findings of weight (or feature) transfusion experiments, we introduce network trimming to further optimize the proposed lightweight All-ConvNet+TL model. This involves selectively pruning the network's weights, leading to the development of a more efficient Lightweight All-ConvNet-Slim model [67].
7. Proposed Histogram of Oriented Gradients (HoG) as a distinctive features characterization method [22] for unique representation of instantaneous HD-sEMG images, adopting pairwise SVMs as the classification scheme. To the author's knowledge, this is the first study to propose HoG as a unique and discriminative feature for sEMG-based gesture recognition.
8. More extensive experiments are conducted. A performance evaluation on four (4) publicly available HD-sEMG datasets was performed on three different sEMG-based gesture recognition tasks: *intra-session*, *inter-session*, and *inter-subject* scenarios. The results showed that the proposed methods outperformed the more complex state-of-the-art models on various tasks and datasets.

## 1.6 Organization of the thesis

The rest of the thesis is organized as follows:

Chapter 2 provides a literature review on sEMG-based gesture recognition. It defines the state-of-the-art methods and identifies the research problems and major drawbacks associated with the current state-of-the-art solutions. Furthermore, it describes the database employed to evaluate the proposed methods and benchmark them against the current state-of-the-art approaches.

Chapter 3 presents S-ConvNet [25], [63] and its underlying design principles. It also presents a domain adaptation method with shallow convolutional neural networks [100] and compares it with the state-of-the-art methods.

Chapter 4 introduces the lightweight All-ConvNet [27] and its design principles. It presents a transfer learning framework, leveraging the lightweight All-ConvNet (All-ConvNet+TL) [67], and compares its performance to state-of-the-art methods.

Chapter 5 presents the proposed Histogram of Oriented Gradients (HoG) feature extraction method [22] with pairwise SVMs as the classification scheme. It discusses experimental results and analyzes the performance of HoG as distinctive features for sEMG-based gesture recognition.

Chapter 6 provides conclusive remarks on the key findings, accomplishments, limitations, and identifies future research directions.

## **Chapitre 2 - HD-sEMG-Based Gesture Recognition – State-of-the-Art**

Gesture recognition based on high-density surface electromyography (HD-sEMG) facilitates the non-invasive analysis and modeling of sEMG signals in both the temporal and spatial domains, opening up new avenues for the next-generation muscle computer interfaces (MCIs). Researchers have focused on developing high-accuracy gesture recognition algorithms. Various feature extraction and classification algorithms based on traditional machine learning methods have been proposed. In recent years, deep learning methods have been proposed for feature learning and to achieve high recognition accuracy. However, the availability of large datasets, power consumption, data memory, and computational speed are important constraints in this context. Researchers also address a more challenging problem of gesture recognition in inter-session and inter-subject scenarios, where different intrinsic and extrinsic factors present formidable challenges. To address this problem, domain-adaptation methods have been proposed for domain invariant feature representation. In addition to summarizing the current research and state-of-the-art on HD-sEMG-based gesture recognition, the research problems and major drawbacks in the existing state-of-the-art solutions are identified. The aims of this chapter are to give an overview of the current research and state-of-the-art on HD-sEMG-based gesture recognition.

## 2.1 Background

A Muscle Computer Interface (MCI), also known as a Human-Computer Interface (HCI), is a technology that allows individuals to interact with computer systems or devices using their neuromuscular activity. It enables users to control and communicate with technology by detecting and interpreting the myoelectrical signals generated by muscles during muscle contraction and relaxation. A muscle is comprised of numerous motor units (MUs), with each MU consisting of a motor neuron and the muscle fibers it innervates. When a motor unit is activated or fires, it generates a distinct electrical signal called motor unit action potential (MUAP). The MUAP represents the cumulative effect of the individual contributions made by the muscle fibers within the motor unit. In essence, the MUAP is the combined electrical signal produced by the firing of all the muscle fibers associated with a specific motor unit. Surface electromyography (sEMG) records the muscle's electrical activity from the skin's surface, providing insights into the generation and propagation of motor unit action potentials (MUAPs) [68], [69]. Fig. 2.1 illustrates the generation and decomposition mechanism of sEMG signals. As physiological signals are closely associated with human motion, surface electromyography (sEMG) signals play a crucial role in human-computer interaction and serve as essential control signals in human-computer interaction systems (HCISs) [11], [70]. sEMG-based gesture recognition has been the technical core of non-intrusive human-computer interfaces (HCI), which are often directed at controlling active prostheses [1], wheelchairs [2], [72], exoskeletons [3], [71], or neurorehabilitation [4], neuromuscular diagnosis [5] and providing an alternative interaction method for video games [6], [7].

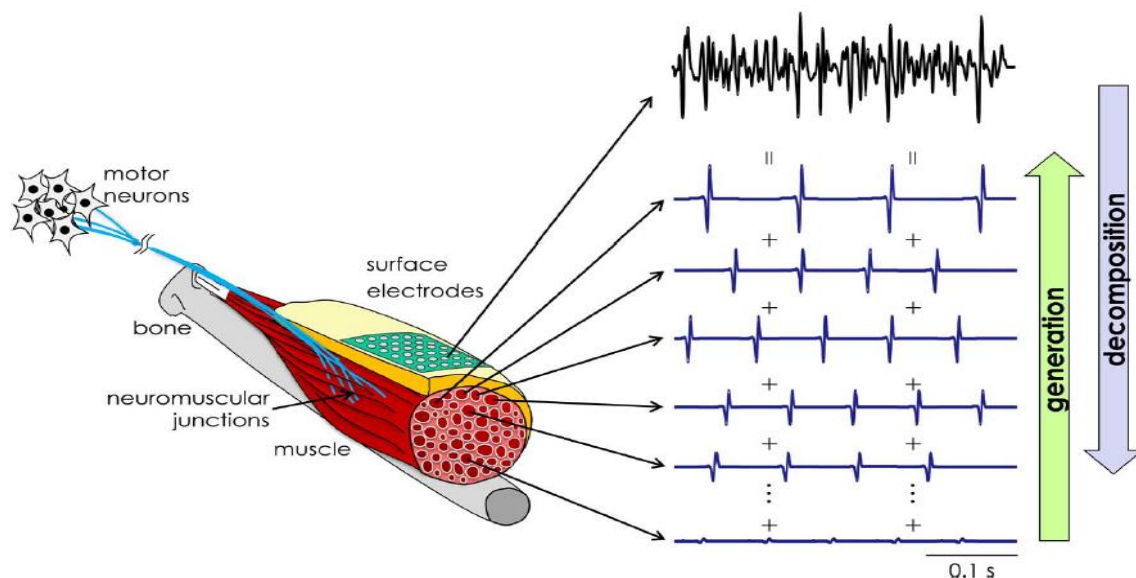


Fig. 2.1 Generation and decomposition mechanism of sEMG signals. Adapted from [96].

The existing sEMG-based gesture recognition methods can be categorized into sparse multi-channel sEMG and high-density sEMG (HD-sEMG) approaches. In sparse multi-channel sEMG-based approaches, the continuous sequences of sEMG signals are usually divided into specific time windows and assigned gesture or muscular activity labels to each window. Afterward, different time-domain, frequency-domain and/or time-frequency domain features [73], [74] are extracted from each of these windowed sEMG signals [11-16]. These windowed features, along with their assigned gesture labels, are used to train an appropriate classifier (e.g., LDA – linear discriminant analysis) [8], [76], SVM – support vector machines [9], HMM – hidden Markov model [10] and MLP – multi-layer perceptron [75]) to classify the new incoming sequential features into various neuromuscular activities or gesture classes during the evaluation or testing phase. A schematic diagram of sEMG-based gesture recognition using conventional sparse multi-channel and windowed sEMG signals is shown in Fig. 2.2. For the evaluation of these developed methods for sparse multi-channel sEMG-based gesture recognition, the most widely accepted benchmark

database is NinaPro [15]. This database involves 67 subjects performing 52 different hand, finger, and wrist gestures, creating two separate databases (DB1 and DB2). However, the reported state-of-the-art recognition accuracy (75.32%) achieved by NinaPro DB1 [15] is not feasible for real-time muscle-computer interface (MCI) applications. Moreover, the sparse multi-channel sEMG-based methods are not suitable for real-world applications due to their limitations related to electrode shift and positioning. Hence, if any malfunction occurs with any one of these sparse multi-channel sEMG electrodes, the entire MCIs system would need to be retrained [17][18]. Deep learning-based methods have recently been explored for gesture recognition using sparse multi-channel sEMG [19-20], [31-32], [61], but their performance is still far from optimal [64].

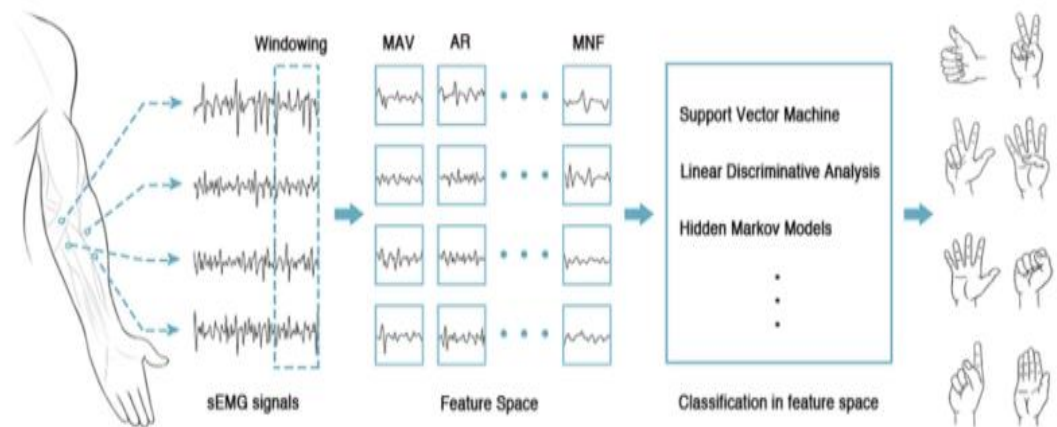


Fig. 2.2 Schematic illustration of sEMG-based gesture recognition by windowing sEMG signals. Adapted from [21]).

To overcome the problem of sparse multi-channel sEMG based approaches, the HD-sEMG-based methods have been proposed in recent years [17], [18], [97]. The HD-sEMG records myoelectric signals using two-dimensional (2D) electrode arrays that characterize the spatial distribution of myoelectric activity over the muscles that reside within the electrode pick-up area [17], [18], [97]. The collected HD-sEMG data are spatially correlated which



enables both temporal and spatial changes and is robust against malfunction of the channels with respect to the sparse multi-channel sEMG data [18]. The HD-sEMG-based methods have been used in biomedical applications for many years [77-79]. In addition to its use in biomedical applications, the HD-sEMG-based methods have recently been used for sEMG-based gesture recognition and for the proportional control of the multiple degrees of freedom (DOFs) for muscle-computer interfaces (MCIs) [17-18], [21-30], [36], [80-81]. In this chapter, we concentrate only on the gesture recognition methods based on HD-sEMG approach since it is the current focus.

The rest of this chapter is organized as follows. Section 2.2 provides an overview of feature extraction and classical/traditional machine learning methods used for HD-sEMG-based gesture recognition. Section 2.3 discusses current state-of-the-art deep learning methods explored for HD-sEMG-based gesture recognition. Section 2.4 presents state-of-the-art methods on inter-session and inter-subject scenarios. Section 2.5 provides an overview of the state-of-the-art CapgMyo HD-sEMG dataset. Finally, Section 2.6 provides some conclusive remarks.

## **2.2 Feature Extraction and Classical Machine Learning (ML) Methods**

The conventional framework for gesture recognition using sEMG involves several key stages: data preprocessing, feature extraction, feature selection, and gesture classification [24]. Within this process, the feature extraction and gesture classification stages hold particular significance in the context of sEMG-based gesture recognition. sEMG signals are characterized by robust nonlinearity. To obtain unique attributes from these sEMG signals, it is typically imperative to extract discriminative features from the pre-processed sEMG signals within a specific time window [11],[82]. Therefore, conventional methods for

sEMG-based gesture recognition focused on extracting distinctive feature sets with domain knowledge. Currently, the following three primary methods are predominantly employed in the literature for sEMG feature extraction: time-domain feature, frequency-domain feature, and time-frequency-domain feature [73-74], [83]. Finding effective hand-crafted features that can discriminate sEMG signals corresponding to different neuromuscular activities or hand gestures is a challenging task. Researchers have spent decades exploring ways to capture both the temporal and frequency information of the sEMG signals. For a comprehensive overview of typical sEMG features and their applications, refer to [11], [74]. These time, frequency and time-frequency features can be summarized as follows:

- **Time-domain features**-includes the root mean square (RMS), integrated EMG (iEMG), zero-crossing (ZC) points, waveform length (WL), variance (VAR), and mean absolute value (MAV). Among these various features considered, RMS and iEMG have been extensively utilized for feature extraction due to their ability to not only capture amplitude variations of sEMG signals in the time domain but also effectively reveal the biomechanical performance of muscles and the fluctuations in muscle energy during motion/movement. The RMS and iEMG features can be computed as follows [87]:

$$RMS = \left( \frac{1}{T} \int_t^{t+T} x^2(t) dt \right)^{\frac{1}{2}} \quad (2.1)$$

$$iEMG = \int_t^{t+T} |x(t)| dt \quad (2.2)$$

where  $x(t)$  represents sampling voltage at time  $t$  and  $T$  represents the sampling time.

- **Frequency-domain features**—includes power spectrum (PS), average frequency (MNF), intermediate frequency (MDF), frequency ratio (FR), cepstrum coefficients (CC), and autoregressive coefficient (AR). To quantitatively describe the power spectrum characteristics of sEMG signals based on the fast Fourier transform (FFT), the median frequency (MF) and the mean power frequency (MPF) are commonly utilized. They are computed as follows [87]:

$$\int_{f_1}^{F_{median}} PS(f).df = \int_{F_{median}}^{f_2} PS(f).df \quad (2.3)$$

$$F_{mean} = \frac{\int_{f_1}^{f_2} f.PS(f).df}{\int_{f_1}^{f_2} PS(f).df} \quad (2.4)$$

where  $PS(f)$  represents the power spectrum of sEMG signals obtained through the FFT, while  $f_1$  and  $f_2$  denote the lowest and highest frequencies of the sEMG band.

- **Time-frequency domain features**— include discrete wavelet transform coefficients (DWTC), discrete wavelet packet transform coefficients (DWPTC), continuous wavelet transform coefficients (CWTC) [88-90].

The above-mentioned features or combinations of these feature sets have been employed as inputs for classical machine learning methods which include (kNN) [92], LDA [8, 76], SVM [9, 16], HMM [10, 91], and MLP [75] for sEMG-based gesture recognition [61].

The following notable/significant works from the literature have exploited hand-crafted features with classical machine learning methods for HD-sEMG-based gesture recognition. Rojas *et al.* [80] utilized three electrode arrays, each with 128 channels. However, an average of 350 channels were used to record sEMG signals for each subject, aiming to cover the entire muscles of interest [93]. The in-house data for four different gestures

corresponding to three effort levels were acquired from the 12 subjects. In their approach, an HD-sEMG map is defined as the time-averaged 2D intensity map of HD-sEMG signals, in which each pixel is the root mean square (RMS) value of a certain channel in a time window (e.g., 500 ms). Then, the time-averaged 2D intensity map is divided into  $3 \times 3$  grids. The different feature sets such as the center of gravity, mean and maximum intensities, the coordinates of the maximum are used to train a LDA classifier for gesture recognition. The average gesture recognition accuracy, ranging from 90% to 98%, is achieved when recognizing four different hand gestures corresponding to three effort levels. Nougrou *et al.*, [34] used two  $8 \times 8$  HD-sEMG sensors (64 channel sEMG) at the posterior and anterior sides of the right forearm of a healthy subject near the elbow to discriminate between 9 different gesture classes based on the wrist movements. They also defined an HD-sEMG map as the time-averaged 2D intensity map of HD-sEMG signals, as described in reference [80]. Then, the time-averaged 2D intensity map of HD-sEMG is divided into  $3 \times 3$  sub-images. The different feature sets such as the center of gravity, the mean amplitude and the percentage of influence from each of these sub-images are extracted. These feature sets were used to train an LDA classifier for gesture recognition. The gesture recognition accuracy of 99.23% is achieved on in-house data collected for a single subject. Amma *et al.*, [17], used an electrode array with 192 electrodes to record a high-density EMG of the upper forearm muscles. The RMS is computed over all windows of length 73.2 ms for each channel of the gesture segment and used as a feature to quantify the muscle activity. The RMS value of an electromyogram shows a strong correlation with the muscle force generated, making it a suitable and commonly employed feature for measuring muscle activity. A baseline system for HD-sEMG-based gesture recognition was introduced, employing a naive Bayes classifier to differentiate among the 27 gestures performed by 5

subjects over five consecutive sessions. An average intra-session gesture recognition accuracy of approximately 90% is achieved. Stango et. al., [18] used 8×24 HD-sEMG sensors (192 channel sEMG) to discriminate 9 different hand gestures performed by seven able-bodied subjects and one unilateral trans-radial traumatic amputee. The spatial correlation of high-density EMG signals was computed on a defined time window (e.g., 50 ms) using a variogram analysis, which measures the spatial variance information by computing the distance between EMG signals in a spatial grid and using them as a unique feature. Fig. 2.3 demonstrates the computation of spatial correlation features of HD-sEMG through variogram analysis. These extracted spatial correlations of HD-sEMG features were used to train a SVM classifier with a linear kernel for gesture recognition. An average intra-session gesture recognition accuracy of approximately 95% is achieved on in-house data for the recognition of nine different hand gestures obtained from seven able-bodied subjects. This research highlights that the spatial correlation computed through variogram analysis offers features that are more robust against noisy channels and electrode shift and positioning compared to the widely used RMS, time domain (TD), and time-domain autoregressive (TDAR) features in the literature for EMG pattern recognition. Khushaba *et al.* [84] proposed a framework for temporal-spatial descriptors (TSD) that involves extracting various time-domain features from the EMG signal within each channel over a specific time window, such as signal energy, spectral moments, zero crossings (ZC), number of peaks (NP), coefficients of variation (COV), and Teager-Kaiser energy operator. These time-domain features are then fused with spatial correlation features, computed by either computing the difference or summation of EMG signals between channels. The resulting fused TSD features are used to train LDA, kNN, and SVM classifiers respectively for gesture classification. The validation of the proposed TSDs was conducted using four

sparse-channel sEMG and an HD-sEMG dataset obtained from both able-bodied individuals and amputees, encompassing a wide range of hand and finger movements. The intra-session gesture classification results revealed significant error rates decreases compared to alternative methods that relied solely on time-domain features, exhibiting an average improvement of at least 8% across all subjects for sparse-channel EMG. Furthermore, the TSDs demonstrated satisfactory performance in HD-sEMG-based gesture recognition tasks.

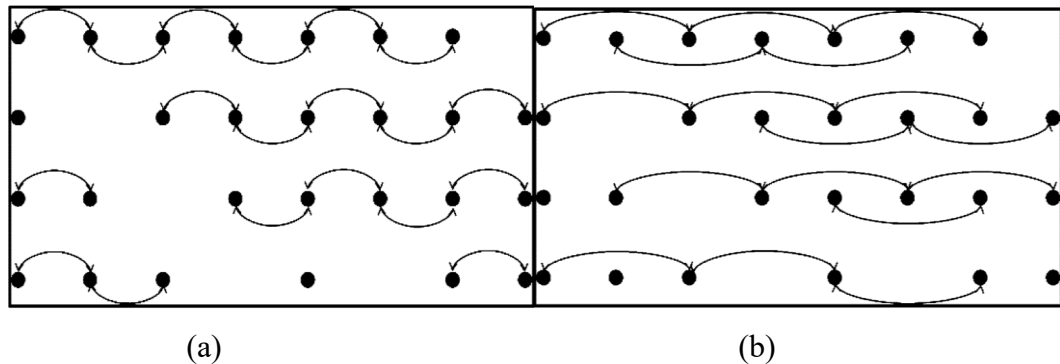


Fig. 2.3 Measurements of all pairs of distances at a) 1-point and b) 2-points apart along the  $x$ -directions for an HD-sEMG grid. Adapted from [18]

However, the existing HD-sEMG-based gesture recognition methods depend on extracting features from windowed sEMG data, necessitating the search for an optimal window length. Determining this optimal window length presents a substantial trade-off between classification accuracy and controller delay, both of which increase with an increase in window size and can significantly influence classification accuracy and controller delay. This is particularly critical in applying assistive technology, physical rehabilitation, and human-computer interfaces [21], [26]. To address this problem and enable the development of more fluid and natural muscle-computer interfaces (MCIs) methodologies, a new feature extraction method based on Histogram of Oriented Gradients (HoG) [22] is introduced in

this thesis for sEMG-based gesture recognition. This method, presented in chapter 5, effectively characterizes and discriminates hand gestures solely captured by instantaneous HD-sEMG signals without needing windowed sEMG signals.

Nevertheless, the classical machine learning methods often require strong reliance on domain-specific knowledge for tasks such as feature extraction, feature selection, and parameter tuning. The appeal of deep learning for sEMG-based gesture recognition lies in its ability to address these challenges by directly incorporating feature learning into the algorithm training process. The following section discusses the current state-of-the-art deep learning methods that have been explored for HD-sEMG-based gesture recognition.

### **2.3 Deep Learning Methods for sEMG-based Gesture Recognition**

Over the past few years, there has been a gradual transition in sEMG-based gesture recognition, moving away from conventional machine learning approaches and towards the adoption of deep learning methods. Deep learning (DL) represents a category of machine learning algorithms that distinguish themselves from traditional ML methods by incorporating feature extraction as an integral part of the model definition, eliminating the requirement for manually engineered hand-crafted features described in the previous section. In the domain of gesture recognition using sparse multi-channel sEMG, there has been a recent exploration of deep learning-based methods [19-20], [31-32], [61]. However, their performance still remains suboptimal [64]. To address this problem, designing and developing more flexible, convenient, and comfortable HD-sEMG based myoelectric sensors and efficient deep learning-based pattern recognition methods have been major research directions in recent years [21], [23-27], [29], [36], [57], [63], [67].

Geng *et al.* [21] present a deep learning methodology for gesture recognition using instantaneous sEMG images. The instantaneous sEMG images were spatially composed of HD-sEMG signals. Then, a deep convolutional neural network (CNN or ConvNet) was employed that resembles DeepFace [35] and trained with these spatially composed instantaneous HD-sEMG images for gesture recognition. Experiments were conducted on publicly available benchmark HD-sEMG datasets [15], [17], [26], and the results demonstrated state-of-the-art recognition accuracy. Fig. 2.4 illustrates the ConvNet architecture employed by Geng *et al.* [21]. Du *et al.*, [26] utilize Adaptive Batch Normalization (AdaBN) [37] to enhance the scalability of the classifier used in [21] for sEMG-based gesture recognition.

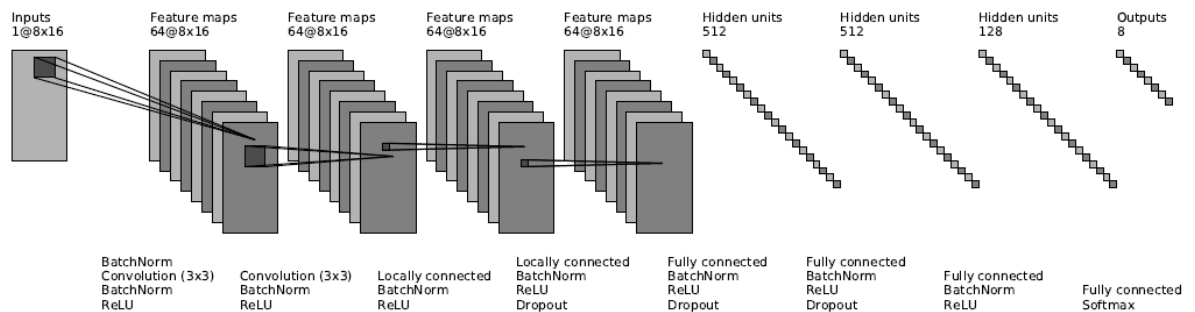


Fig. 2.4 A schematic illustration of the ConvNet architecture employed by Geng *et al.* [21]. Adapted from [26].

Wei *et al.* [23] proposed a two-stage CNN with a multi-stream decomposition stage and a fusion stage to learn the correlation between certain muscles and specific gestures. The sEMG image is decomposed into different equally sized image patches based on the layout of the electrode arrays on muscles (e.g., each of eight  $8 \times 2$  electrode arrays in the CapgMyo database [26] individually produces  $8 \times 2$  equal-sized sEMG image patches). Then, each of these sEMG image patches is independently and in parallel passed through the convolution layers of a single-stream CNN [21], thereby forming a multi-stream CNN. The learned



features from all the single-stream CNNs that form a multi-stream CNN are aggregated and fed to a fusion network for gesture recognition. The reported results showed that multi-stream CNN outperformed single-stream CNN by a small margin. Fig. 2.5 illustrates the multi-stream decomposition and fusion network.

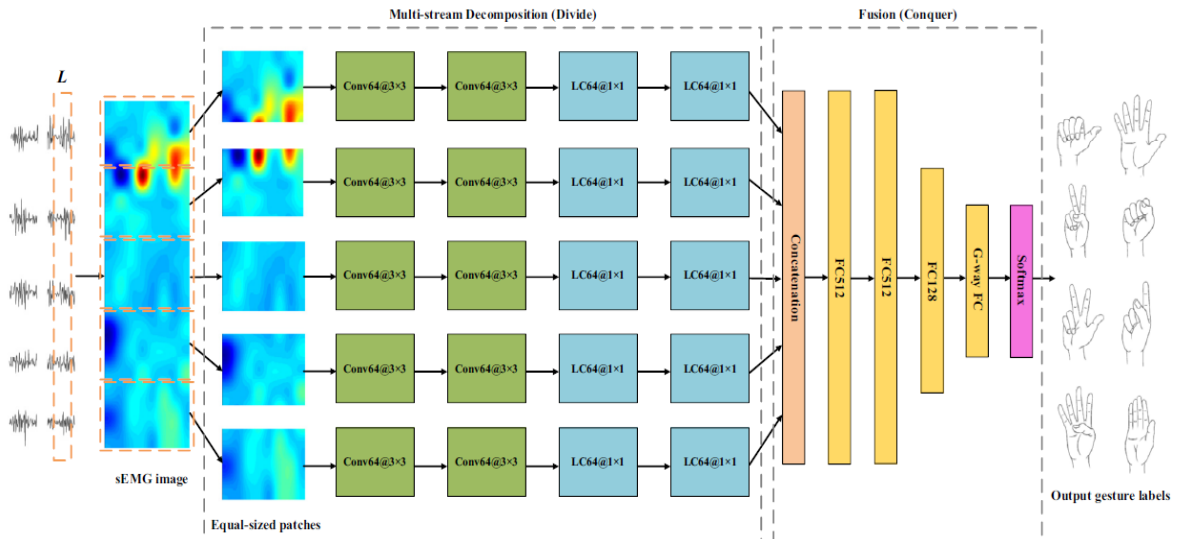


Fig. 2.5 A schematic illustration of multi-stream decomposition and fusion network. Adapted from [23].

Hu *et al.* [24] proposed a combination of CNN and recurrent neural network (RNN) called CNN-RNN network with an attention module to capture both spatial and temporal information of sEMG signals for gesture recognition. The recorded sEMG signals were decomposed into small subsegments using a sliding and overlapping windowing strategy. Each of these sEMG subsegments was converted into an sEMG image and simultaneously passed through a multi-stream CNN built upon [21] for feature extraction. Given the input sequence of the extracted features corresponding to each of the sEMG subsegments, a long short-term memory (LSTM) network was learned individually for gesture recognition. Then, the features learned by each of these LSTMs corresponding to each of these sEMG

subsegments were concatenated before being fed to a fully connected and SoftMax layer for gesture recognition. Experimental results indicate that a combined CNN-RNN network with an attention module outperforms the stand-alone CNN and RNN frameworks, respectively. Fig 2.6 elaborates on the CNN-RNN network with an attention mechanism.

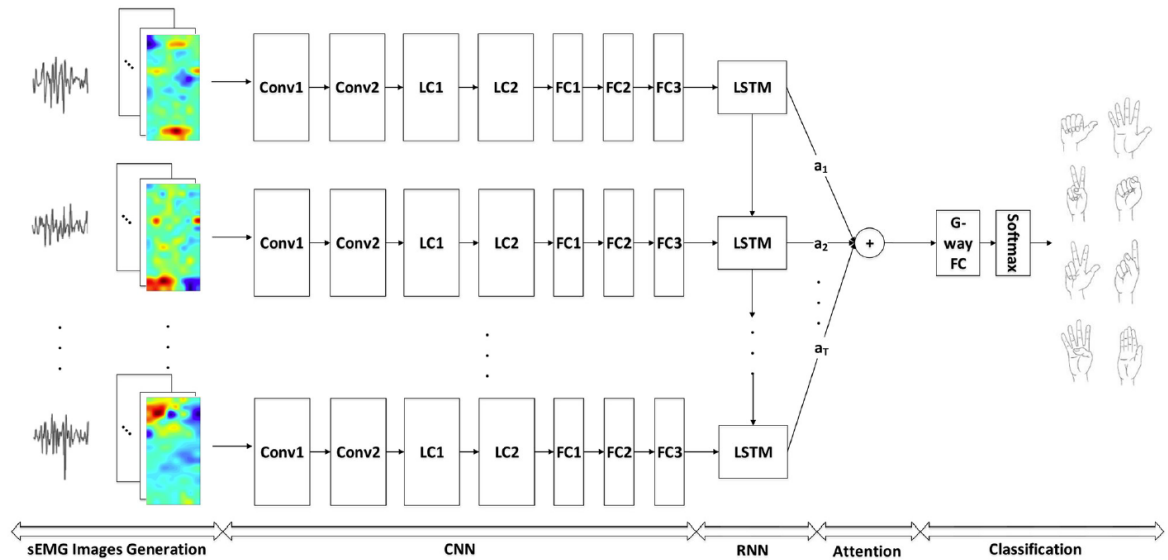


Fig. 2.6 A schematic elaboration of CNN-RNN network with an attention module for sEMG-based gesture recognition. Adapted from [24].

Encouraged by [38], Chen et al. proposed using 3D convolution in the convolutional layers of CNNs for spatial and temporal representation of sEMG images [36]. The 3D convolution is attained by convolving a 3D kernel to the cube formed by stacking multiple adjacent sEMG image frames. The feature maps in the convolution layers of a 3D CNN are connected to multiple adjacent sEMG image frames in the previous layer. Hence, the spatiotemporal information is captured. However, multiple 3D convolutions with distinct kernels are required to apply at the same location of the input to learn representative features, which makes 3D CNN computationally expensive. For example, the exploited 3D CNN in [36] requires learning over >30M (million) parameters when the length of the input

cube is set to 10 (i.e., the cube is formed by stacking 10 consecutive sEMG image frames). Fig. 2.7 describes the 3D convolutional neural network architecture employed by Chen et al., [36] for sEMG-based gesture recognition.

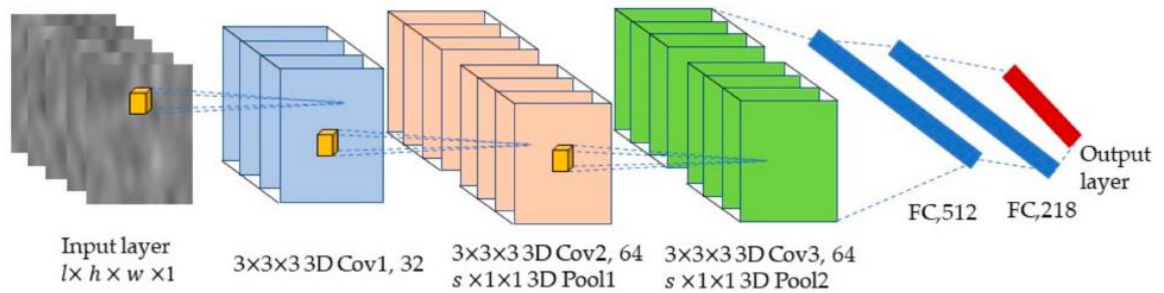


Fig. 2.7 A schematic diagram of the 3D convolutional neural network architecture. Adapted from [36].

However, the state-of-the-art methods [21], [23], [24] for sEMG-based gesture recognition either employed very complex deep and wide CNN or an ensemble of these complex networks for improved gesture recognition performance. For example, Geng *et al.* [21] exploited a DeepFace [35] like very large and deep CNN (dubbed as GengNet), which requires learning  $>5.63\text{M}$  (million) training parameters only during fine-tuning and pre-trained on a very large-scale labelled sEMG training datasets. The complexity of this model grows linearly as the input size is increased due to the use of an unshared weight strategy [27]. Wei *et al.* [23] used an ensemble of eight (8) single-stream GengNet at the decomposition stage only. Hu et al. [24], used a two-stage ensemble network in which an ensemble of multiple single-stream GengNet was used for spatial feature learning, resulting in multiple sequences of 1-D feature representation. Then, these 1-D feature sequences were passed to an ensemble of LSTM networks before a SoftMax layer recognized the targeted gesture. Despite the significant performance boost achieved by these state-of-the-art models [21], [23], [24], the high computational and intensive memory cost hinders

deploying them on resource-constrained embedded and mobile devices for real-time applications. Therefore, the demand for designing low-cost, lightweight networks is highly increasing for low-end resource-limited embedded and mobile devices.

The next section presents several factors that bottleneck sEMG-based gesture recognition, especially in inter-session and inter-subject scenarios and outlines the current state-of-the-art solutions to address this problem.

## **2.4 Inter-Session/Inter-Subject Scenarios**

The sEMG-based gesture recognition approaches discussed in the previous section are usually investigated in intra-session scenarios because this is currently regarded as the gold standard in sEMG-based gesture recognition [17]. However, real-time human-computer interfaces (HCIs) based on gesture recognition using sEMG signals suffer from various intrinsic and extrinsic factors that result in data variability between the source domain task and the target domain task. In sEMG-based gesture recognition, the term 'source domain task' refers to the original task or sEMG dataset from which the model is trained. Conversely, the 'target domain task' involves deploying the model, trained on the source domain task, to a different task (e.g., new recording sessions or an unseen subject or dataset) that may originate from a different context or environment and can be significantly impacted by various intrinsic and extrinsic factors. The intrinsic factors refer to the elements that influence the generation of the EMG signal. These factors include individual differences, muscle fatigue, variations in contraction force, contraction patterns, and other related aspects. Extrinsic factors pertain to the elements that impact the process of collecting the EMG signal. These factors include variations in electrode placement, electrode-skin contact impedance, variations in downstream task requirements, etc. [94].

Due to these intrinsic and extrinsic factors that deviate the source domain task from the target domain task, the sEMG-based gesture recognition in inter-session and inter-subject scenarios becomes a highly challenging task.

- **Inter-Session**-in real-time applications, the models are usually built by training with the data collected from the previous sessions and deployed to new sessions for MCI using sEMG-based gesture recognition. However, the data distribution in these new sessions may differ due to various factors, such as variations in how gestures are performed, differences (shifts) in electrode placement, channel variations, and variations in muscle contraction force or muscle fatigue. *Inter-session* refers to the scenario when a classifier is trained using data obtained from participants during one session, and its performance is subsequently evaluated using data recorded during a separate session [26], [57], [63], [67], [94].
- **Inter-subject**-EMG signals exhibit significant variation among individuals, which can be attributed to differences in subcutaneous fat distribution, muscle fiber diameter, and the different way of performing force/gestures. *Inter-subject* refers to the scenarios when a classifier is trained using data from a group of subjects, and its performance is evaluated using data from an unseen subject who was not included in the training data [26], [57], [63], [67], [94].

To mitigate the effects of electrode shift and displacements, Hargrove *et al.*, [95] first identified all possible electrode displacement locations during the HD-sEMG data acquisition process. Then, the classifier was trained with the data acquired from all these electrode displacement locations to recognize gestures. Amma *et al.* [17] employed the Gaussian mixture model (GMM) along with a small amount of calibration data to address

the challenges posed by electrode shifts and displacements between two sessions. Implementing these techniques led to a significant improvement in the inter-session recognition accuracy of 27 gestures, increasing it from 58.9% to 75.4%. Khushaba [124] introduced a feature transformation approach based on canonical correlation analysis (CCA) aimed at reducing data variation in sEMG-based gesture recognition. Patricia et al. [16] present multi-source adaptive learning algorithms that leverage the Geodesic Flow Kernel with an SVM classifier. The experiments were conducted on the NinaPro dataset, which consists of sparse channel sEMG data, and achieved an inter-subject recognition accuracy of approximately 40% for 52 gestures [26]. Moreover, to address these intrinsic and extrinsic factors that cause inter-session and inter-subject data variability, currently [26] and [57] provide state-of-the-art solutions. Du *et al.* [26] proposed a multi-source extension to the classical adaptive batch normalization (AdaBN) technique [37] for domain adaptation, specifically designed to be used with CNN architecture. The drawback of this solution is that when dealing with multiple sources (i.e., multiple subjects), it is necessary to impose specific constraints and considerations for each source during the pre-training phase of that model [57]. Ketyko *et al.* [57] from *Nokia Bell Labs*, proposed a 2-Stage recurrent neural network (2SRNN), where a deep stacked RNN sequence classifier was used for pre-training on the source sEMG dataset. Then, the weights of the pre-trained deep-stacked RNN classifier were frozen. At the same time, a fully connected layer without a non-linear activation function was trained in a supervised manner on the target dataset for domain adaptation. More explicitly, the deep-stacked RNN classifier was used as a feature extractor by freezing its weight in the domain adaptation stage. However, ConvNet is more powerful at extracting discriminative features than RNN, even for classification tasks of long sequences [58], [59].

To cope with the distribution shift problem, the current state-of-the-art methods [26], [57] employed very large and complex deep ConvNet or 2SRNN-based domain adaptation methods to approximate the distribution shift caused by these inter-session and inter-subject data variability due to the above-mentioned intrinsic and extrinsic factors. Hence, these methods require learning over millions of training parameters and a large pre-trained and target domain dataset in both the pre-training and adaptation stages. Therefore, deploying these high-end, resource-constrained, and computationally expensive methods becomes challenging for real-time applications. Hence, designing and developing efficient, lightweight domain-invariant feature representation methods are highly demanded for sEMG-based gesture recognition.

## 2.5 CapgMyo Dataset

Currently CapgMyo [26] and CSL-HD-sEMG [17] are the two most used HD-sEMG datasets for the evaluation of HD-sEMG-based gesture recognition methods. The CSL-HD-sEMG dataset [17] was collected only from five (5) subjects. Whereas the CapgMyo dataset [26] were collected from twenty-three (23) subjects and divided into three sub datasets in order to evaluate sEMG-based gesture recognition methods in intra-session, inter-session and inter-subject scenarios. Therefore, the CapgMyo datasets were used to evaluate, and technical validation of the proposed sEMG-based gesture recognition methods discussed in the subsequent chapters. In addition, the CapgMyo database was adopted in this research because this is the first database that were developed and made publicly available especially for evaluating the performances of gesture recognition methods based on instantaneous HD-sEMG images. This dataset was developed with the objective of providing a standard benchmark database to explore the new possibilities for

studying next-generation muscle computer interfaces (MCIs). The acquisition device utilized in constructing the CapgMyo dataset consists of 8 EMG acquisition modules, each equipped with a matrix-type ( $8 \times 2$ ) array of differential electrodes, resulting in a combined total of ( $8 \times 16$ ) 128 sEMG channels. The sEMG acquisition modules were affixed to the subject's right forearm using adhesive bands. The first acquisition module was positioned on the extensor digitorum communis muscle, aligned with the radio-humeral joint, while the subsequent sEMG modules were evenly distributed in a clockwise direction from the subject's viewpoint. Moreover, the CapgMyo database comprises 3 sub-databases (referred to as DB-a, DB-b and DB-c). In DB-a, a total of 18 out of 23 subjects participated in the data acquisition process while performing 8 isotonic and isometric hand gestures. Each gesture in DB-a was held for a duration of 3 to 10 seconds and repeated 10 times, with the EMG signals sampled at a sampling frequency of 1000 Hz. DB-b consists of the same gesture set as DB-a but was acquired from 10 out of the 23 subjects. Each gesture in DB-b was held for approximately three (3) seconds. Every subject in DB-b contributed two recording sessions executed on different days. It's worth highlighting that the recording interval between the two sessions in Db-b of the CapgMyo dataset is greater than one week, and the placement of the electrode array varies at each recording session. Consequently, this benchmark dataset can also provide valuable insights into the effectiveness of various approaches under the scenario of electrodes variation [94]. In DB-c, 12 basic finger movements were obtained from 10 out of the 23 subjects. Each gesture in DB-c was also held for approximately three (3) seconds. It should be noted that the gesture set used in CapgMyo dataset was a subset of the NinaPro [15] database, with the shared objective of encompassing most of the finger movements encountered in activities of daily living. Additionally, this subset facilitated a comparison of the gesture recognition performance



based on HD-sEMG and sparse multi-channel sEMG signals. Tables 2.1 and 2.2 illustrate gestures in CapgMyo DB-a, DB-b and DB-c respectively.

Table 2.1 Gestures in DB-a and DB-b (8 isotonic and isometric hand configurations). Adapted from [26].





















Label	Description	Instance	Label	Description	Instance
1	Thumb up		5	Abduction of all fingers	
2	Extension of index and middle, flexion of the others		6	Fingers flexed together in fist	
3	Flexion of ring and little finger, extension of the others		7	Pointing index	
4	Thumb opposing base of little finger		8	Adduction of extended fingers	

Table 2.2 Gestures in DB-c (8 isotonic and isometric hand configurations). Adapted from [26].

Label	Description	Instance	Label	Description	Instance
1	Index flexion		7	Little finger flexion	
2	Index extension		8	Little finger extension	
3	Middle flexion		9	Thumb adduction	
4	Middle extension		10	Thumb abduction	
5	Ring flexion		11	Thumb flexion	
6	Ring extension		12	Thumb extension	

## 2.6 Conclusion

In this chapter, an overview of the current sEMG-based gesture recognition methods, including feature extraction and classical machine learning methods, current state-of-the-art deep learning methods, state-of the-art HD-sEMG datasets and the related issues including different factors that impede HD-sEMG-based gesture recognition in inter-session and inter-subject scenarios and its existing state-of-the-art solutions are discussed.

Current state-of-the-art deep learning methods for gesture recognition based on instantaneous HD-sEMG signals are high-end resource intensive, therefore not feasible for deploying in real-time MCI applications due to on-device constraints of computational speed, data memory, power consumption and processing of large datasets. To overcome these problems, the next chapter introduces low-latency and parameter efficient S-ConvNet, along with a domain adaptation method leveraging S-ConvNet, for discriminative and domain-invariant feature representation for improved gesture recognition based on instantaneous HD-sEMG signals in intra-session, inter-session, and inter-subject scenarios.

## **Chapitre 3 - Domain Adaptation with Low-Latency Shallow Convolutional Neural Networks for Improved Inter-Session/Inter-Subject Gesture Recognition**

The concept of sEMG based gesture recognition using instantaneous HD-sEMG images and underlying deep representation learning opens up new avenues for the development of more fluid and natural muscle-computer interfaces. However, the existing approaches employed a very large and complex deep ConvNet architecture and complex training schemes for HD-sEMG image recognition, which requires learning of  $>5.63$  million(M) training parameters only during fine-tuning and pre-trained on a very large-scale labeled HD-sEMG training dataset. As a result, it makes high-end resource-bounded and computationally very expensive for deployment in real-time applications. To overcome this problem, S-ConvNet models are proposed, a simple yet efficient framework for learning instantaneous HD-sEMG images from scratch using random-initialization. Without using any pre-trained models, S-ConvNet proposed demonstrate and set a new state-of-the-art performance while reducing learning parameters to only  $\approx 2$ M and using  $\approx 12 \times$  smaller dataset for sEMG-based gesture recognition in intra-session scenarios. Moreover, the distribution shift is a challenging problem for gesture recognition in inter-session and inter-subject scenarios. To further address this challenging problem, a domain adaptation method with a shallow CNN is proposed. The proposed domain adaptation method with S-

ConvNet outperformed the current state-of-the-art methods based on a very large and complex deep ConvNet or 2-stage Recurrent Neural Networks (2SRNN) by a large margin both when the data from single trials or multiple trials are available for domain adaptation. Experiments conducted on four (4) publicly available HD-sEMG datasets, described in Chapter 2, on different sEMG-based gesture recognition tasks such as intra-session, inter-session and inter-subject scenarios validate the effectiveness of the proposed methods. The state-of-the-art performance on various HD-sEMG datasets and tasks proved that the proposed methods are highly effective for learning discriminative and domain-invariant representations for instantaneous HD-sEMG image recognition, especially in the data and high-end resource-constrained scenarios.

### **3.1 Introduction**

Gesture recognition based on instantaneous HD-sEMG signals is usually employed in portable, mobile, and wearable device applications within the context of real-time Muscle-Computer Interfaces (MCIs). Hence, there is a high demand for designing and developing low-latency shallow CNN architectures to enable on-device inference on these mobile wearable devices while maintaining state-of-the-art accuracy. However, the current state-of-the-art methods [21], [23], [24], [26] and [61] employed a DeepFace [35] like very large ConvNet architecture for gesture recognition based on sEMG signals, which requires to be pre-trained on a very large-scale labeled training dataset even for gesture recognition in intra-session/within session scenarios. For example, in [23], a multi-stream extension of the DeepFace [35] like ConvNet, as employed in [21], [26], is proposed. An ensemble of multi-stream CNN built upon [21] and long-short term memory networks (LSTM) integrated with an attention module is proposed in [24]. In addition, Chen et. al. [36] used a 3D CNN for

learning spatial and temporal representation of sEMG images, however this model requires learning  $> 30 M$  parameters for instantaneous HD-sEMG image recognition. As a result, these state-of-the-art network models becomes only high-end resource bounded and computationally very expensive, making them impractical for real-world MCIs applications.

Apart from that, the network architectures employed by the current state-of-the-art approaches [21], [23], [26], [61] are heavily rely on pre-training with large-scale HD-sEMG training datasets ( $\approx 0.76$  million), even for sEMG-based gesture recognition in intra-session scenarios. The conventional paradigm of using pre-trained models in the literature when the *source task A* is different from the *target task B* and when there are not enough target data available to make the training accomplishable alone on the *target task B* [39]. However, in the existing approaches for instantaneous HD-sEMG image recognition (e.g., [21], [23], [26] and [61]), both the *source task A* and the *target task B* are the same, and pre-training on large-scale HD-sEMG training datasets has been performed with the aim of preventing overfitting during re-training or fine-tuning using the data available for the *target task B* i.e., intra-subject and intra-session test. Therefore, it is not surprising to achieve high target task accuracy with these highly resource-based and fined-tuned network architectures for gesture recognition in intra-session scenarios. Also, this conventional wisdom of pre-training is recently challenged by He *et al.* [39], where pre-training does not necessarily improve the target task accuracy is proved to be claimed. Hence, we hypothesize that the requirement of initializing the target network using the pre-trained weights for *intra-session* gesture recognition is due to the large and complex deep models employed by these current state-of-the-art.

Following are the other critical limitations of using pre-trained networks for instantaneous HD-sEMG image recognition:

- (i) *Constrained structure design space* – pre-trained networks employed by [21], [23], [26] and [61] are very deep and large and trained on a large-scale HD-sEMG dataset, therefore, containing a massive number of parameters. Hence, there is a little flexibility to control/adjust the network structures (even for small changes) by directly adopting the pre-trained network to the target task. The requirement of computing resources and large-scale pre-trained datasets are also bounded by large network structures [25], [63], [99].
- (ii) *Domain mismatch* – the existing sEMG based gesture/neuromuscular activity recognition methods are usually trained and evaluated on the data acquired from the able-bodied subjects. However, in real time sEMG-based MCIs applications (e.g., assistive technology, physical rehabilitation etc.) are most of the time designed for elderly people, amputees and patients. These differences impose a serious problem due to the varied sEMG distributions in the *source* and *target tasks*. Though the fine-tuning of the pre-trained model can reduce the gap, however, it is still a serious problem, when there is a huge mismatch between the source and the target task [25], [63],[98].
- (iii) *Learning bias* – the distributions and the loss functions between the *source task* and the *target task* may vary significantly, which may lead to different searching/optimization spaces. Therefore, the learning may be biased towards a local minimum which is not optimal for the target task [25], [63],[99].

However, it is legitimate to adopt a pre-trained network for sEMG-based gesture recognition in *inter-session* and *inter-subject* scenarios (considered as different tasks), where the distribution of the sEMG signals may be different/shifted from the sEMG signals used to train the model in the previous sessions or to an unseen new subject.

To overcome these above-mentioned problems, this thesis work is motivated by the following research question- *is it possible to train the sEMG-based gesture recognition model from scratch without any pre-training while still maintaining state-of-the-art performance?* To achieve this goal, we propose a shallow convolutional neural network (S-ConvNet) [25], [63] architecture, a simple yet effective framework, which could learn neuromuscular activity from scratch using only the makeshift HD-sEMG dataset available for the target task/subject. The S-ConvNet is reasonably flexible and can be tailored to various network structures for different computing platforms, such as desktops, servers, mobile devices, and even embedded electronics. Despite its simplicity and flexibility, the S-ConvNet achieves state-of-the-art performance across different sEMG-based gesture recognition tasks, including intra-session, inter-session, and inter-subject scenarios. Moreover, it outperforms the more complex current state-of-the-art methods while significantly reducing the number of learning parameters.

Furthermore, with the equivalent or competitive accuracy to that of current state-of-the-art methods, S-ConvNet with fewer parameters has the following advantages [123]: (1) S-ConvNet requires less communication across servers during distributed training. (2) S-ConvNet require less bandwidth to export a new model from the cloud to mobile or wearable edge devices for executing a target gesture recognition task and on-device

inference. (3) S-ConvNet is more feasible to deploy on FPGAs and other hardware with limited memory.

For instantaneous sEMG image-based gesture recognition, the challenge remains open because very limited research has been done on it. The main contributions of this chapter are summarized as follows.

- (1) We present S-ConvNet: A shallow convolutional neural network architecture [25], [63], according to the best of my knowledge, this is the first ConvNet framework that enables training instantaneous HD-sEMG based gesture recognition model from scratch without any pre-training and achieve state-of-the-art performance outperforming the most complex current state-of-the-art methods.
- (2) Propose a domain adaptation method with low-latency shallow convolutional neural networks [100] to approximate the domain shift for enhancement of sEMG-based gesture recognition accuracy in inter-session and inter-subject scenarios.
- (3) We perform extensive experiments on four (4) publicly available HD-sEMG datasets: CapgMyo DB-a, DB-b (Session 1), DB-b (Session 2) and DB-c [26] respectively on three different sEMG-based gesture recognition tasks: intra-session, inter-session, and inter-subject scenarios. The proposed S-ConvNet achieves superior performance for intra-session gesture recognition when trained from scratch on the target sEMG dataset and improves inter-session and inter-subject gesture recognition accuracy through domain adaptation.

This chapter is organized as follows: Section 3.2 presents the proposed S-ConvNet framework, while Section 3.3 provides the model description, design principles and



training methodology for S-ConvNet, respectively. Section 3.4 presents experimental results and discusses the performance of the proposed S-ConvNet models and compared with the current state-of-the-art models for instantaneous HD-sEMG based gesture recognition in intra-session scenarios. Section 3.5 introduces domain adaptation methods with low-latency shallow convolutional neural network (S-ConvNet) for the enhancement of inter-session and inter-subject gesture recognition accuracy. Section 3.6 evaluates sEMG-based gesture recognition in inter-session and inter-subject scenarios and compares it against state-of-the-art methods. Section 3.7 offers some conclusive remarks.

### **3.2 The Proposed Framework**

The proposed framework for sEMG-based gesture recognition using instantaneous HD-sEMG images includes the following three major computational components: (i) pre-processing and sEMG image generation (ii) architectural design of the S-ConvNet model and (iii) classification. A schematic diagram of the proposed framework of sEMG-based gesture or neuromuscular activity recognition by instantaneous sEMG images is shown in Fig. 3.1. First, the power-line interferences were removed from the acquired HD-sEMG signals with a band-stop filtered between 45 and 55 Hz using a 2nd order Butterworth filter. Then, the HD-sEMG signals at each sampling instant were arranged in a 2-D grid according to their electrode positioning. This grid was further transformed into an instantaneous sEMG image by linearly transforming the values of sEMG signals from mV to color intensity as  $[-2.5\text{mV}, 2.5\text{mV}]$  to  $[0\ 255]$ . Thus, an instantaneous grayscale sEMG image was formed with the size of  $16 \times 8$ . Secondly, we devised different S ConvNet models which describe in Section 3.3. Finally, providing instantaneous HD-sEMG images and their corresponding gesture labels, the devised S-ConvNet model is trained offline to predict to

which gesture or muscular activity an instantaneous HD-sEMG image belongs. Then, this trained S-ConvNet model is used to recognize different gestures or neuromuscular activities at test time from the unseen instantaneous HD-sEMG images.

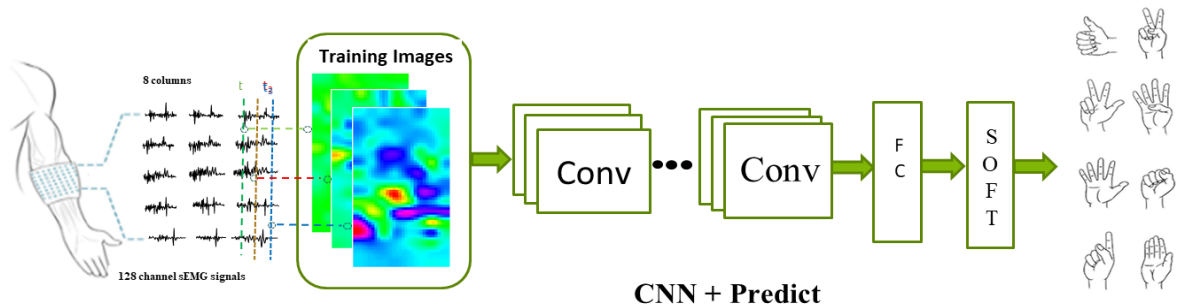


Fig. 3.1 Schematic diagram of the proposed framework of muscular activity recognition by instantaneous sEMG images.

### 3.3 Model Description– The Shallow Convolutional Neural Network (S-ConvNet)

S-ConvNet network architectures differ from existing approaches for HD-sEMG image recognition in several key aspects. Firstly, S-ConvNet models are trained from random initialization i.e., from scratch without any pre-training for gesture recognition in intra-session scenarios. The pre-training in the existing approaches (e.g. [21], [26] and [23], [61]) involves over 720k images acquired from 18 different subjects. However, considering the targeted application domains of sEMG-based gesture recognition (e.g., assistive technology, physical rehabilitation etc.), it is always difficult to gather such a large amount dataset required for the pre-training and fine-tuning of very large and complex deep neural network (DNN) models. We cannot expect an amputee or a patient to provide a large set of training examples over a multiple number of trials and sessions. Thus, acquiring such high quality abundant labeled data are often limited, expensive and inaccessible in the domain of sEMG analysis for training these complex state-of-the-art deep ConvNet models [94].

Moreover, there is currently no evidence if this specialized very large and complex DNN architecture is either required or needs to be pre-trained with large scale sEMG dataset for instantaneous HD-sEMG image recognition in intra-session scenarios. This work demonstrates that it is possible to attain state-of-the-art accuracy by the proposed simple yet efficient S-ConvNet network model architectures when being trained from scratch directly on target sEMG dataset for intra-session gesture recognition, outperforming the highly resource-intensive, pre-trained, and fined-tuned current state-of-the-art [21], [26], [23] network model architectures. Training from random initialization, S-ConvNet models require  $\approx 12 \times$  *smaller dataset* than its pre-trained counterparts for HD-sEMG image recognition. Fig. 3. 2 shows the total number of images used during training for pre-training + fine-tuning vs random initialization.

Fig. 3.2 Total number HD-sEMG images seen during training, for pre-training + fine-tuning vs. random-initialization.



Secondly, the network architecture of the existing methods for HD-sEMG image recognition requires pre-training using a large-scale HD-sEMG dataset. Therefore, the question arises of which components of CNNs are necessary for achieving competitive performance as per these existing methods from random initialization. Motivated by the work in [46], a first step towards answering this question is taken by studying the simplest architecture conceivable: a network consisting of convolution layers, with a maximum of one fully connected layer with a small number of neurons and an occasional dimensionality reduction by using a *max/average pooling* or using a stridden CNN. The

use of a small number of convolutions and fully connected layers in S-ConvNet greatly reduces the number of parameters and thus also serves as a form of regularization.

Thirdly, the HD-sEMG image classifier requires normalization to help the optimization. In addition to deploying successful forms of normalized parameter initialization methods [52], [53], employing an effective activation normalization method is equally important when training an instantaneous sEMG image recognition model such as S-ConvNet from scratch. Batch Normalization (BN) [55] is a widely used activation normalization technique in the development of deep learning-based methods. BN is used to normalize features by computing the mean and variances over mini-batches of instantaneous HD-sEMG image samples, which has also shown promise in many other applications for easing optimization and enabling deep networks to converge faster. Moreover, the stochastic uncertainty of the batch statistics provides some form of regularization which may yield better generalization [101]. In addition to BN, Dropout [56] is another most popular regularization technique and a simple way to prevent deep neural networks from overfitting. However, Dropout and BN often lead to worse performance when they are combined. This is due to the fact that the Dropout would shift the variance of a specific neural unit when the state of the network transfer from training to test. On the other hand, BN would maintain its statistical variance, which is accumulated from the entire training process, in the test phase. These inconsistencies in variance cause unstable numerical behavior when the signal goes deeper through the network, which may even lead to incorrect predictions [102]. Unlike the existing approaches, Dropout and BN applied separately in an initial experiment with the proposed S-ConvNet models and evaluated their respective performance.

### 3.3.1 S-ConvNet Architecture and Training

The proposed S-ConvNet was trained on a multi-class sEMG-based gesture or neuromuscular activity recognition task to recognize a neuromuscular activity class through an instantaneous HD-sEMG image. The overall architecture of S-ConvNet models is described in Table 3.1. Starting from the simplest Model A, the depth and number of parameters in the network gradually increases to Model C. The instantaneous HD-sEMG image is passed through a convolutional (conv.) layer, where a small receptive field with a  $3 \times 3$  filters are used. The smallest receptive field with  $3 \times 3$  filters is the minimum filter size to allow overlapping convolutions and spatial pooling with a stride of 2, which also capture the notion of left, right and center amicably.

Table 3.1 The three S-Convnet networks Models for sEMG-based gesture recognition using instantaneous sEMG images.

Model A	Model B	Model C
<b>Input <math>16 \times 8</math> Gray-level Image</b>		
$3 \times 3$ Conv. 32 ELU	$3 \times 3$ Conv. 32 ELU	$3 \times 3$ Conv. 32 ELU
$3 \times 3$ Conv. 64 ELU	$1 \times 1$ Conv. 32 ELU	$3 \times 3$ Conv. 32 ELU
$3 \times 3$ Conv. 64 ELU	$3 \times 3$ Conv. 64 ELU	$3 \times 3$ Conv. 32 ELU with stride $r = 2$
FC1 256 ELU	$1 \times 1$ Conv. 64 ELU	$3 \times 3$ Conv. 64 ELU
FC2 G-way FC	FC1 256 ELU	$3 \times 3$ Conv. 64 ELU
and/or	FC2 G-way FC	$3 \times 3$ Conv. 64 ELU with stride $r = 2$
FC3 G-way SoftMax	and/or	FC1 256 ELU
-	FC3 G-way SoftMax	FC2 G-way FC
-	-	and/or
-	-	FC3 G-way SoftMax

It can be observed that the Model B from Table 3.1 is a variant of the Network in Network architecture [47], where only  $1 \times 1$  convolution is performed after each normal  $3 \times 3$  convolutions layers. The  $1 \times 1$  convolution acts as a linear transformation of the input

channels followed by a non-linearity [103]. We also highlight that the model C is a variant of the simple ConvNet models introduced by J. T. Springenberg et. al., [46] for object recognition in which the spatial pooling is performed by using a stridden CNN. The operation of a convolution map and a subsequent spatial pooling are illustrated in Fig. 3.3. The output of a convolution map  $f$  produced by a convolution layer  $c$  is computed as follows:

$$c_{i,j,o}(f) = \Phi \left( \sum_{h=1}^k \sum_{w=1}^k \sum_{u=1}^n \theta_{h,w,u,o} \cdot f_{g(h,w,i,j,u)} \right) \quad (3.1)$$

where  $\theta$  are the convolutional weights or filters;  $g(h, w, i, j, u) = (r \cdot i + h, r \cdot j + w, u)$  is the function mapping from position in  $c$  to position in  $f$  respecting the stride  $r$ ;  $w$  and  $h$  are respectively the width and height of the filters;  $n$  is the number of channels (in case  $f$  is the output of a convolutional layer,  $n$  is the number of filters);  $o \in [1, M]$  is the number of output feature or channels of the convolutional layer and  $\Phi(\cdot)$  is the activation function, an exponential linear unit ELU defined as:

$$\Phi(x) = \begin{cases} \alpha(\exp(x) - 1), & \text{if } x < 0 \\ x, & \text{if } x \geq 0 \end{cases} \quad (3.2)$$

Then, the spatial pooling operations are performed to the convolution map  $f$  as follows:

$$s_{i,j,u}(f) = \left( \sum_{h=1}^k \sum_{w=1}^k |f_{g(h,w,i,j,u)}|^p \right)^{\frac{1}{p}} \quad (3.3)$$

The equation (3.3) can be interpreted as a form of *max pooling* when  $p \rightarrow \infty$  [46]. In addition, by making a slight adjustment to equation (3.3) and setting  $p = 1$ , the following equation can be translated into an *average-pooling*:

$$s_{i,j,u}(f) = \left( \frac{1}{k} \sum_{h=1}^k \sum_{w=1}^k |f_{g(h,w,i,j,u)}|^p \right)^{\frac{1}{p}} \quad (3.4)$$

The spatial pooling operation on a convolutional map makes the networks more robust to local translations and may help cope with the electrode shifting problem encountered in different HD-sEMG recording trials and sessions. However, the spatial pooling operations may cause the network to lose distinctive information about the detailed texture and micro-textures of an instantaneous sEMG image. Therefore, the pooling operations are only introduced to our models after the first convolutional layers in order to investigate the effect of these pooling operations on our network models.

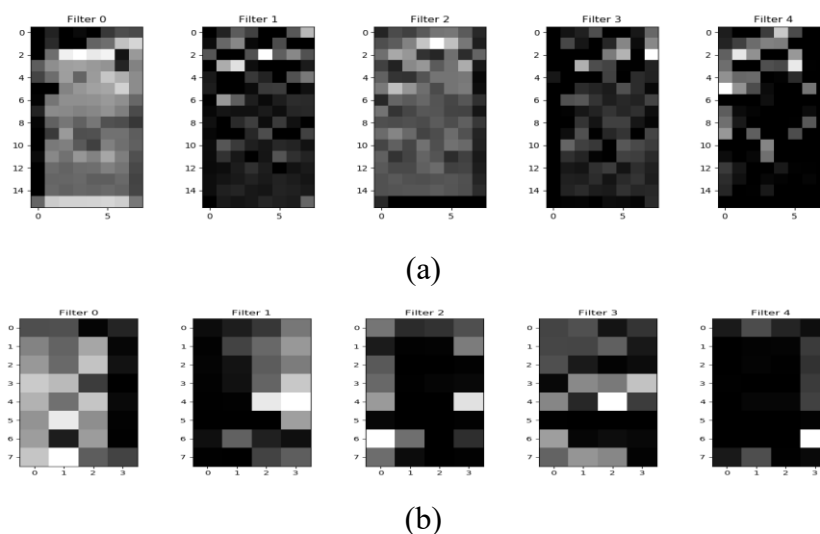


Fig. 3.3 A schematic illustration of convolutions and pooling operation a) Convolution maps and b) Convolutions maps after spatial pooling.

Afterwards, the convolution maps produced by the final convolutional layer of each of the model networks, illustrated in Table 3.1, are flattened out and concatenated to form a multi-dimensional feature vector. Then, the flattened feature vector is inputted to a fully connected layer where each of the feature elements are connected to all its input neurons. This fully connected layer can capture correlations between features extracted in the distant part of the instantaneous sEMG images. The output of the fully connected layer in the network is used as discriminative feature representation for instantaneous sEMG images. In

terms of representation, this is in contrast to the HOG-based sEMG image representation [22], described detailed in Chapter 5, that generally extract very local descriptors by computing the histograms of oriented gradients and use as input to a classifier.

Finally, the output of the fully connected layer is fed to a  $G$ -way SoftMax layer (where  $G$  is the number of hand gesture or neuromuscular activity classes) which produces a distribution over the class labels. If we denote  $\hat{y}^{(j)}$  as the  $j$ th element of the  $G$  dimensional output vector of the layer preceding the SoftMax layer, the class probabilities are estimated using the SoftMax function  $\sigma(\cdot)$  defined as below:

$$\sigma(\hat{y}^{(j)}) = \frac{\exp(\hat{y}^{(j)})}{\sum_G \exp(\hat{y}^{(G)})} \quad (3.5)$$

The goal of this training is to maximize the probability of the correct gesture or neuromuscular activity class. We achieve this by minimizing the cross-entropy loss [49] for each training sample. If  $y$  is the true label for a given input, the loss is

$$L = -\sum_j y^{(j)} \ln(\sigma(\hat{y}^{(j)})) \quad (3.6)$$

The loss is minimized over the parameters by computing the gradient of  $L$  with respect to the parameters and by updating the parameters using the state-of-the-art Adam (adaptive moment estimation) gradient descent-based optimization algorithm [50].

Having trained the network, an instantaneous HD-sEMG image is recognized as in the gesture or neuromuscular activity class  $C$  by simply propagating the input image forward and computing:

$$C = \operatorname{argmax}_j(\hat{y}^{(j)}) \quad (3.7)$$



### 3.3.2 Normalization

As discussed above, the acquired HD-sEMG signals at each sampling instant were arranged in a 2-D grid according to their electrode positioning and converted into an instantaneous sEMG image by linearly transforming the values of sEMG signals from mV to color intensity as  $[-2.5\text{mV}, 2.5\text{mV}]$  to  $[0\ 255]$ . Therefore, the intensity distribution of the transformed sEMG images is normalized to be between zero and one i.e.  $[0\ 1]$  in order to reduce the sensitivity to contrast and illumination changes. Given an input sEMG image  $I$ , this is accomplished by applying *max-min* normalization as follows:

$$I_N = \frac{I - I_{min}}{I_{max} - I_{min}} \quad (3.8)$$

where  $I_{max}$  and  $I_{min}$  are, respectively, the maximum and minimum pixel intensity of the input instantaneous HD-sEMG image  $I$ , and  $I_N$  is the normalized instantaneous HD-sEMG image within the range  $[0, 1]$ . It is worth mentioning that our training data were not pre-normalized when the batch normalization (BN) is applied.

The performance of the proposed S-ConvNet models were evaluated online by learning the instantaneous sEMG image representation on CapgMyo<sup>1</sup> database [26] as described in detail in Chapter 2, Section 2.5, for sEMG-based gesture recognition. Next section discusses experimental results and analysis to evaluate the performance of the proposed S-ConvNet in *intra-session* scenarios as well as some insight and findings about learning and recognizing instantaneous sEMG images.

---

<sup>1</sup>The dataset is made publicly accessible from the following website: [http://zju-capg.org/research\\_en\\_electro\\_capgmyo.html](http://zju-capg.org/research_en_electro_capgmyo.html).

### 3.4 The Performance Evaluation of The Proposed S-ConvNet Network Models

From the viewpoint of MCI application scenarios, sEMG-based gesture recognition can be categorized into three types [26], [63], [57], [67]:

Type I – intra-session: in which a classifier is trained on time part of the data recorded from the subjects during one session and evaluated on another time part of the data recorded from the same session,

Type II – inter-session: in which a classifier is trained on the data recorded from the subjects in one session and tested on the data recorded in another new session, and

Type III – inter-subject: when a classifier is trained using data from a group of subjects, and its performance is evaluated using data from an unseen (new) subject who was not included in the training data.

However, the sEMG-based gesture recognition in the literature has usually been investigated in intra-session scenarios (e.g., [18], [21], [23], [24]) because this is currently regarded as the gold standard in sEMG-based gesture recognition [17]. In this section, we evaluate the performance of the proposed S ConvNet models in intra-session scenarios and compare them with the current state-of-the-art. However, the implications of our proposed methods in inter-session and inter-subject scenarios are reported in Sections 3.5 and 3.6 of this chapter, respectively.

The CapgMyo database [26] comprises three sub-databases (referred to as DB-a, DB-b, and DB-c) with a total of 23 able-bodied subjects aged between 23 and 26 years. Notably, DB-b was recorded in two different sessions, named DB-b (Session 1) and DB-b (Session 2),

respectively. Therefore, all four sub-datasets are considered for intra-session performance evaluations.

In DB-a, we obtained 8 isotonic and isometric hand gestures from 18 out of the 23 subjects, with each gesture recorded 10 times. In DB-b, the same gesture sets as in DB-a were recorded from ten (10) different subjects. Additionally, in DB-c, twelve (12) gestures with each gesture repeated 10 times were recorded from each of the ten (10) different subjects who participated in this sub-database. In all these mentioned CapgMyo datasets, the gestures set were recorded with a 1000 Hz sampling rate. For each subject, the recorded HD-sEMG data is filtered, sampled and instantaneous sEMG image is generated using the method described in Section 3.2. More explicitly, each target subject in DB-a performed 8 different hand gestures and each gesture were recorded 10 times with a 1000 Hz sampling rate, therefore, each target subject participated in DB-a in total generates  $(8 \times 10 \times 1000) = 80\,000$  or 80k instantaneous sEMG images individually.

### 3.4.1 Data Selection for Training, Validation and Testing

Existing approaches (e.g., [21], [26], [23], [61]) for instantaneous HD-sEMG image recognition use a pre-trained model for sEMG-based gesture recognition in intra-session scenarios. In their approaches, the pre-trained dataset consists of combined instantaneous sEMG images obtained from odd-numbered trials performed by each target subject. For illustration, in DB-a, 18 target subjects participated, performing 8 different hand gestures, with each gesture being trialed 10 times. From these 10 trials, only the instantaneous images obtained from the odd-numbered trials from each target subject (i.e.,  $8 \times 5 \times 1000$ ) = 40,000 or 40k were combined, resulting in a total of  $(18 \times 40,000) = 720,000$  or 720k

instantaneous images. These combined training instantaneous sEMG images from all the target subjects were used for pre-training.

The pre-trained model was then used to develop a subject-specific classifier by fine-tuning it with training data obtained exclusively from a specific target subject. Specifically, the pre-trained model underwent fine-tuning using 40,000 (or 40k) training samples consisting of instantaneous sEMG images from the odd-numbered trials. The resulting fine-tuned model was then tested using the remaining 40,000 (or 40k) instantaneous sEMG images from the even-numbered trials of the same target subject (e.g., in DB-a). Thus, the existing approaches [21], [26], [23], [61] involve a total of  $(720,000 + 40,000) = 760,000$  or 760k images in the training process for developing a target subject-specific classifier (as illustrated in Fig. 3.2).

In contrast, the proposed S-ConvNet model is trained from scratch through random initialization without any pre-training for intra-session gesture recognition. The training, validation, and testing process of the proposed model utilizes only the makeshift dataset available for a target subject, i.e.,  $(1 \times 8 \times 10 \times 1000) = 80,000$  or 80K images produced by a target subject individually for DB-a and DB-b, respectively. In DB-c, each of the target subjects produces  $(1 \times 12 \times 10 \times 1000) = 120,000$  or 120K instantaneous sEMG images individually because each subject in DB-c performed 12 hand gestures and each gesture repeated 10 times with a 1000 Hz sampling rate.

In order to maximize the use of instantaneous sEMG images during training, we introduce the Leave-One-Trial-Out Cross-Validation (LOTOV) approach. In LOTOV, we systematically leave out one (1) trial in turn from the pool of 10 different trials representing 8 distinct hand gestures, resulting in 8,000 or 8k instantaneous sEMG images for testing.

The remaining 9 trials, comprising a total of 72k images for 8 different hand gestures, are used for training and validation. From these 72k training images, we randomly select 9k images for validation to assess whether our devised model is prone to overfitting during the training process. Therefore, our training process involves only 63k images, which contrasts with existing approaches that employ 760k images in the training process (as shown in Fig 3.2).

Finally, the leave one trial out cross-validation accuracy for 10 different test trials is averaged and used as a performance indicator for an intra-subject and intra-session test. The cross-validation accuracy  $A$  is computed for each class  $i$ , as the number of totals correctly recognized hand gestures,  $C$ , divided by the total number of tests sEMG images as follows:

$$\text{Accuracy, } A = \frac{C}{N} = \frac{\sum C_i}{N} \quad (3.9)$$

where  $i = \{1, 2, \dots, G\}$  and  $G$  is the number of gesture classes.

### 3.4.2 Experimental Results on Intra-Session Scenarios

In this experiment, we compared all the proposed S-ConvNet models described in Section 3.3 on the CapgMyo and its four (4) sub-datasets without any pre-training or data augmentations. For effective and faster training of a CNN network model with high-dimensional parameter spaces requires a good initialization strategy for the connection weights, a good activation function, using BN and a good regularization technique. The weight initialization scheme, an activation function and the effectiveness of BN and Dropout regularizers were determined in a preliminary experiment using S-ConvNet model

A (see in Table 3.1) and involving subjects 2 from CapgMyo DB-a. Then, these choices were subsequently maintained throughout all subsequent experiments.

We investigate and experiments with two different initialization schemes for the connection weights in Xavier and He initialization [52], [53]. However, we found that the models with He initialization scheme perform on average 1.5% worse than the Xavier initialization counterparts. We hence do not report the results in this chapter with the He initialization to avoid cluttering the experiments. In order to investigate a most suitable activation function for the proposed S-ConvNet models, we also performed experiments with the different activation functions [48], [104]. The results are reported in Table 3.2. In addition, as the BN and Dropout often lead to worse performance when they are combined as discussed in Section 3.3. In order to investigate this claim, we performed experiments with both of these methods combined and separately. The results are detailed in Table 3.2.

Table 3.2 Gesture recognition accuracy (%) using instantaneous HD-sEMG Images for different activation functions and spatial pooling.

Network	Relu	Leaky-Relu	Elu	Sigmoid	Max-pool	Avg-pool	Avg-run time(min.)
Model A	95.18	95.56	93.98	<b>95.76</b>	94.55	94.31	2.55
Model A with BN	96.16	97.34	97.50	<b>98.00</b>	96.66	97.13	7.74
Model A with Dropout regularization	<b>96.99</b>	96.68	96.58	96.30	95.19	95.61	11.27
Model A with BN and Dropout	97.18	97.54	<b>98.29</b>	97.80	96.98	97.54	14

Also, training a CNN network with a high-dimensional parameter spaces requires an efficient optimization algorithm. Objective functions are often stochastic because of internal data sub-sampling, dropout regularization and other sources of noise. Hence, we

propose to use a computationally efficient stochastic optimization algorithm, Adam [50], which requires only first-order gradients with little memory requirement, is invariant to diagonal rescaling of the gradients and is suitable for high-dimensional problems. It also provides fast and reliable learning convergence that can be considerably faster than the stochastic gradient descent (SGD) optimization algorithm used in the existing approaches for instantaneous HD-sEMG image recognition. Our proposed S-ConvNet models were trained using Adam optimization algorithm with momentum decay and scaling decay are initialized to 0.9 and 0.999 respectively. In contrast to SGD, Adam is an adaptive learning rate algorithm therefore, it requires less tuning of the learning rate hyperparameter. The learning rate 0.001 is initialized to all our experiments. The smaller batches of 256 randomly chosen samples from the training dataset are fed to the network during consecutive learning iterations for all our experiments. We set a maximum of 100 epochs for training S-ConvNet network models. However, to avoid overfitting we have also applied early stopping in which the training process is interrupted if no improvements in validation loss are noticed for 5 consecutive epochs. The BN is applied after the input and before each non-linearity. Dropout was used to regularize all networks. The Dropout was applied on all layers with probabilities 35% for all S-ConvNet models, respectively. The results for S-ConvNet Model A that we conducted for subject 2 in CapgMyo DB-a database are presented in Table 3.2.

Several trends can be observed from these results, the major observations are the following:

1. confirming previous results from the literature, the simplest model A (S-ConvNet A) perform remarkably well, achieving a correct gesture recognition

rate of 98.29% for an intra-subject test based on instantaneous (or per-frame) sEMG images.

2. simply applying *max-min* normalization to the training dataset and fed to the S-ConvNet Model A, the average correct recognition rate of 94.89% has been achieved for 6 different experiments with only 2.55 min overall runtime for training, validation and testing on a Nvidia Tesla K-20C GPU.
3. simply replacing the *max-min* normalization by introducing BN to the network the average correct recognition rate improved to 97.13% by sparing overall 7.74 min runtime for training, validation and testing.
4. the correct recognition rate slightly decreases to 96.23% when the BN is replaced by the Dropout regularizer and *max-min* normalization, while also increasing the overall runtime to 11.27 min for entire training, validation and testing process.
5. when BN and Dropout with a tiny probability are respectively applied to all layers of the network, the average recognition rate increases up to 97.56%. However, due to introducing BN and Dropout to the networks, the overall runtime also increases to about 14 min for entire training, validation and testing process.

In all cases, the performance of the model slightly decreases with spatial *max-pooling* and *average-pooling*. However, spatial pooling can help to regularize CNNs and might be more effective especially when conducting experiments in *inter-session* scenarios. It is worth noting that the average pooling performs quite well when the BN or in conjunction of BN and Dropout are introduced to the network model (e.g., Table 3.2). All these preliminary experimental results for an intra-subject and intra-session test confirm that our proposed S-



ConvNet models can learn all the necessary invariances that requires to build a distinctive representation for instantaneous HD-sEMG image recognition.

From Table 3.2, it can be observed that the maximum correct recognition rate of 98.29% is achieved when applying the exponential linear unit (ELU) activation function, BN, and a small dropout probability to every layer of the network. Therefore, this configuration is also applied to other S-ConvNet B and S-ConvNet C network models respectively and performed experiments on all the subjects who participated in CapgMyo DB-a, DB-b (Session 1), DB-b (Session 2) and DB-c respectively. The results for the proposed S-ConvNet A, S-ConvNet B and S-ConvNet C models are respectively presented in Table 3.3 and compared against the current state-of-the-art methods. In Table 3.3, GengNet [21] is considered as the baseline model because of its adoption to other recent studies [26], [23], [24], [61] and its state-of-the-art performance on various sEMG datasets for intra-session sEMG-based gesture recognition.

As can be seen in the Table 3.3, the simple S-ConvNet models (on the order of  $\approx 2M$  learning parameters) trained from random-initialization with  $3 \times 3$  convolutions and a dense layer with only a smaller number of neuron achieve state-of-the-art results for CapgMyo DB-a, DB-b (Session 1) and DB-b (Session 2) respectively and performs comparable to the state-of-the-art for CapgMyo DB-c dataset even though the state-of-the-art model use more complicated network architectures and training schemes which requires to learn over  $\approx 5.63 M$  learning parameters during fine-tuning only and also pre-trained with over 720k instantaneous HD-sEMG images.

Table 3.3 The average recognition accuracies (%) of 8 hand gestures for CapgMyo DB-a and DB-b for 18 and 10 different subjects respectively and 12 gestures for 10 different subjects in DB-c. The numbers are the majority voted results using 160 ms window (i.e., 160 frames). per-frame accuracies are shown in parenthesis.

Model	S-ConvNet A	S-ConvNet B	S-ConvNet C	Geng et. al., [21], [26]
CapgMyo DB-a	<b>98.36 (87.95)</b>	97.97 (87.02)	97.84 (86.99)	98.48 (86.92)
CapgMyo DB-b Session 1	<b>97.87 (83.57)</b>	97.43 (82.29)	97.42 (82.67)	97.04 (81.26)
CapgMyo DB-b Session 2	<b>97.05 (84.73)</b>	96.40 (83.64)	96.62 (83.90)	96.26 (83.21)
CapgMyo DB-c	95.80 (81.63)	94.47 (79.81)	94.23 (80.04)	<b>96.36 (82.23)</b>
#Learning parameters	$\approx 2.09\text{M}$	$\approx 2.14\text{M}$	$\approx 2.22\text{M}$	$\approx 5.63\text{M}$

Among the proposed S-ConvNet models, S-ConvNet A demonstrates the highest gesture recognition accuracy across all four (4) CapgMyo datasets. Furthermore, both S-ConvNet B and S-ConvNet C outperform the state-of-the-art GengNet [21], [26] model on the CapgMyo DB-b (Session 1) and DB-b (Session 2) datasets, respectively. Fig. 3.4 illustrates the per-frame gesture recognition accuracy, based on instantaneous sEMG images, along with its statistical significance obtained by the proposed S-ConvNet models, compared with the current state-of-the-art GengNet model [21], [26], for 18 different subjects in CapgMyo DB-a. The proposed S-ConvNet models show lower standard deviation over the current state-of-the-art GengNet [21], [26] model. Notably, the S-ConvNet A exhibits superior performance and lower standard deviation among the compared models. A similar performance graph was achieved for CapgMyo DB-b (Session 1 and Session 2) and CapgMyo DB-c, respectively. However, they are not presented here in order to avoid

cluttering the experimental results. Overall, these experimental results validate the stability and potentiality of the proposed S-ConvNet models over the current state-of-the-art. Moreover, these experimental results also indicate that the state-of-the-art GengNet model [21], [26] requires pre-training with a large-scale HD-sEMG dataset as well as fine-tuning on the target dataset for enhancing intra-session gesture recognition accuracy due to its deep and complex large network architecture.

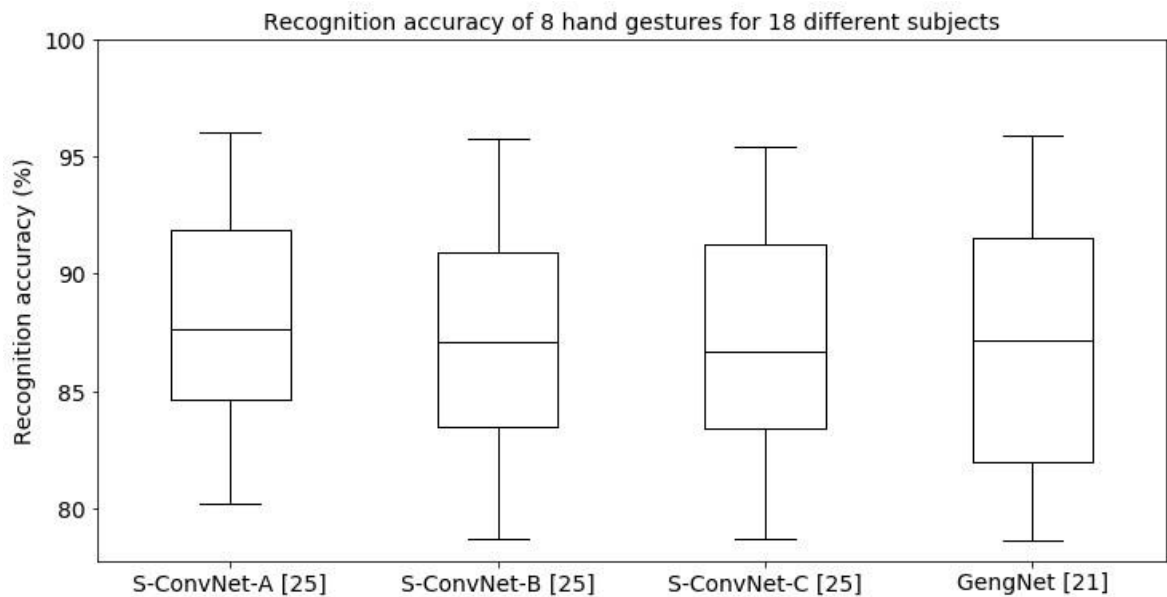
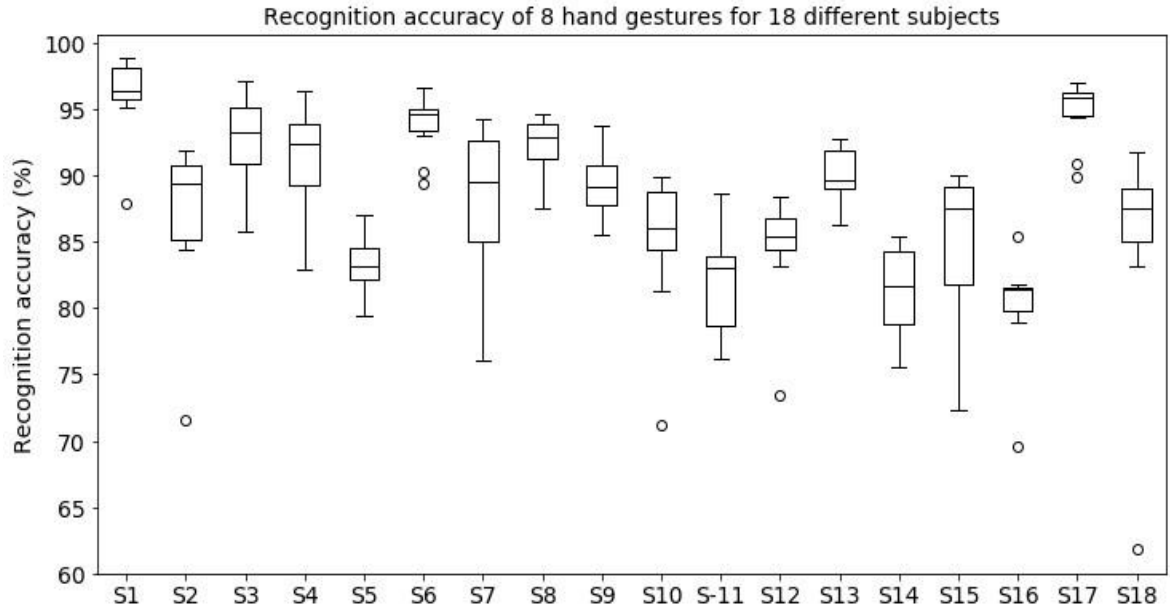
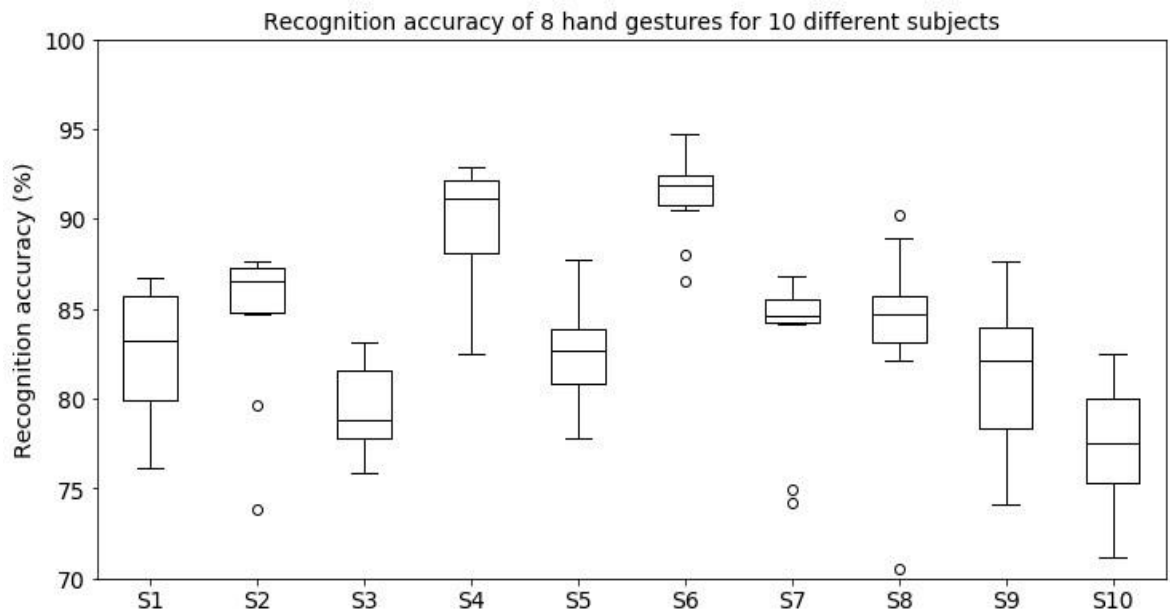


Fig. 3.4 The average per-frame gesture recognition accuracy of 8 hand gestures for 18 different subjects in CapgMyo DB-a with the proposed S-ConvNet models and the current state-of-the-art GengNet [21], [26] model.

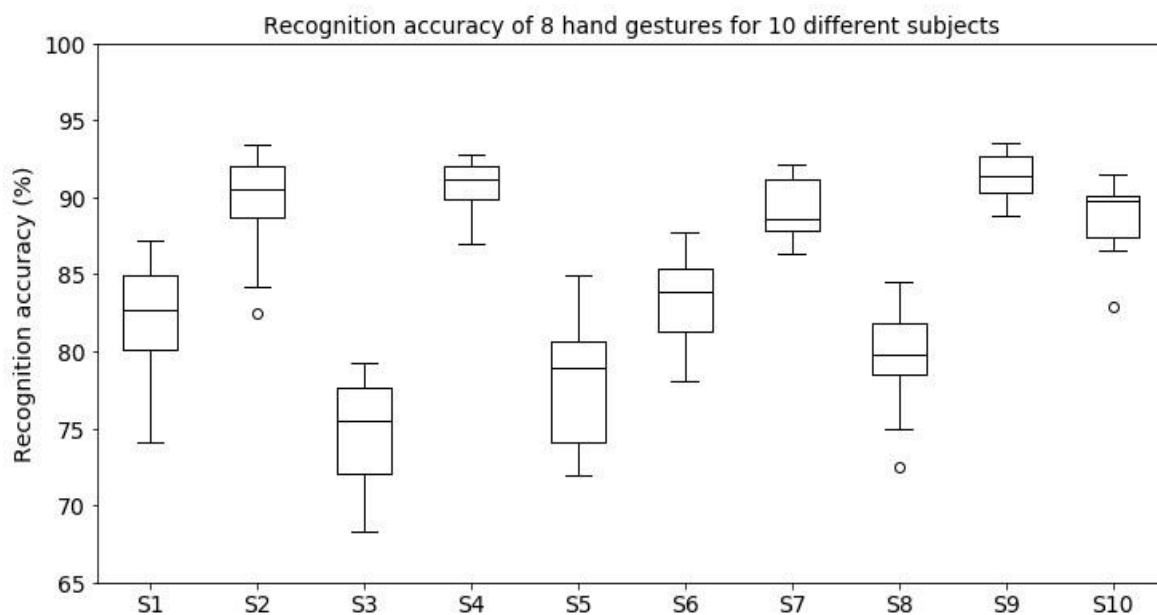
Fig. 3.5 (a)-(d) presents the sEMG-based instantaneous (or per-frame) gesture recognition accuracies and their statistical significance obtained by S-ConvNet A through leave-one-trial-out cross-validation for ten different test trials for each of the participating subjects in CapgMyo DB-a, DB-b (Session 1 and Session 2), and DB-c, respectively. The highest instantaneous (or per-frame) gesture recognition accuracies of 87.95% for DB-a, 83.57% and 84.73% for DB-b (Session 1 and Session 2, respectively), and 81.63% for DB-c, which were obtained by the proposed best performant model S-ConvNet A. These high per-frame



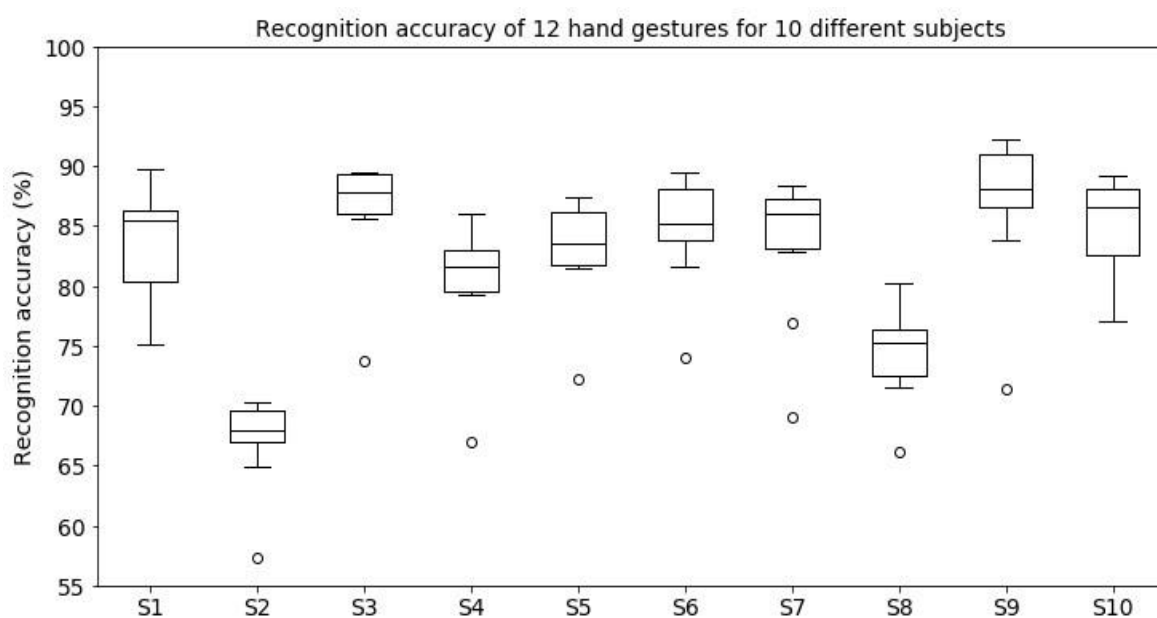
(a) CapgMyo DB-a.



(b) CapgMyo DB-b (Session 1).



(c) CapgMyo DB-b (Session 2).



(d) CapgMyo DB-c.

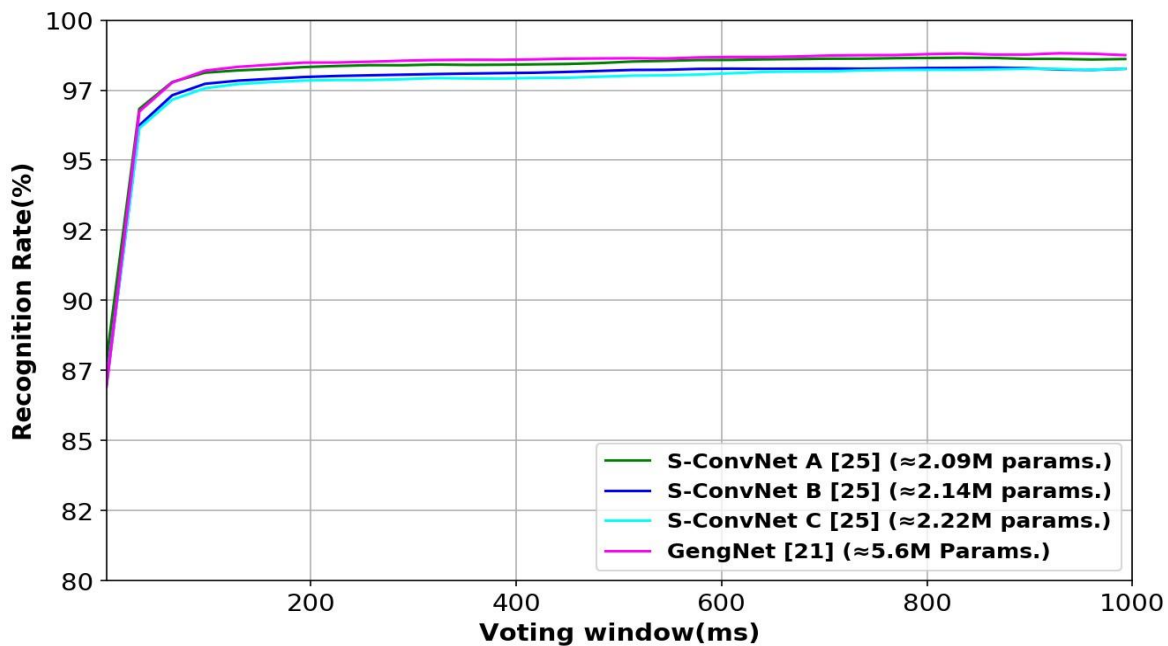
Fig. 3.5 The per-frame gesture recognition accuracy with the proposed S-ConvNet A (a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, (b)-(c) The gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo DB-b (Session 1) and DB-b (Session 2) respectively (d) the gesture recognition accuracy of 12 hand gestures for 10 different subjects on CapgMyo DB-c.

gesture recognition accuracies and low standard deviation over multiple test trials and subjects in each of the above-mentioned four HD-sEMG datasets reflect the high stability of the proposed S-ConvNet network models.

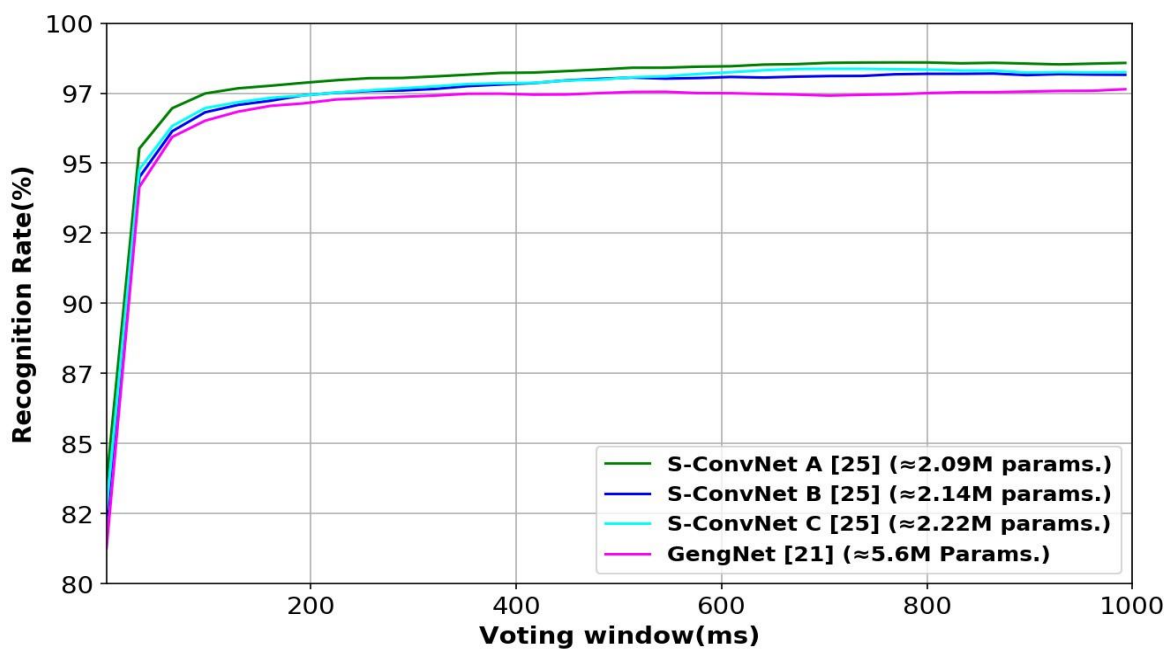
Furthermore, we achieved very good gesture recognition accuracies using a simple majority voting algorithm. Fig. 3.6 (a)-(d) illustrates gesture recognition accuracy with different voting windows using S-ConvNet models and compared against the current state-of-the-art. The average gesture recognition accuracy of 95.02%, 96.31% and 97.01% were achieved by S-ConvNet A with applying a simple majority voting of 32, 64 and 128 instantaneous images (or frames) for the abovementioned four (4) HD-sEMG datasets.

The higher gesture recognition accuracies of 98.36%, 97.87%, 97.50%, and 95.80% (as shown in Table 3.3 and Fig. 3.6) can be obtained by the proposed S-ConvNet A and a simple majority voting over the recognition result of 160 frames for DB-a, DB-b (Session 1 and Session 2) and DB-c, respectively. These results highlights that the proposed S-ConvNet models outperform the current state-of-the-art GengNet [21], [26] at least in three (3) out of four (4) HD-sEMG datasets.

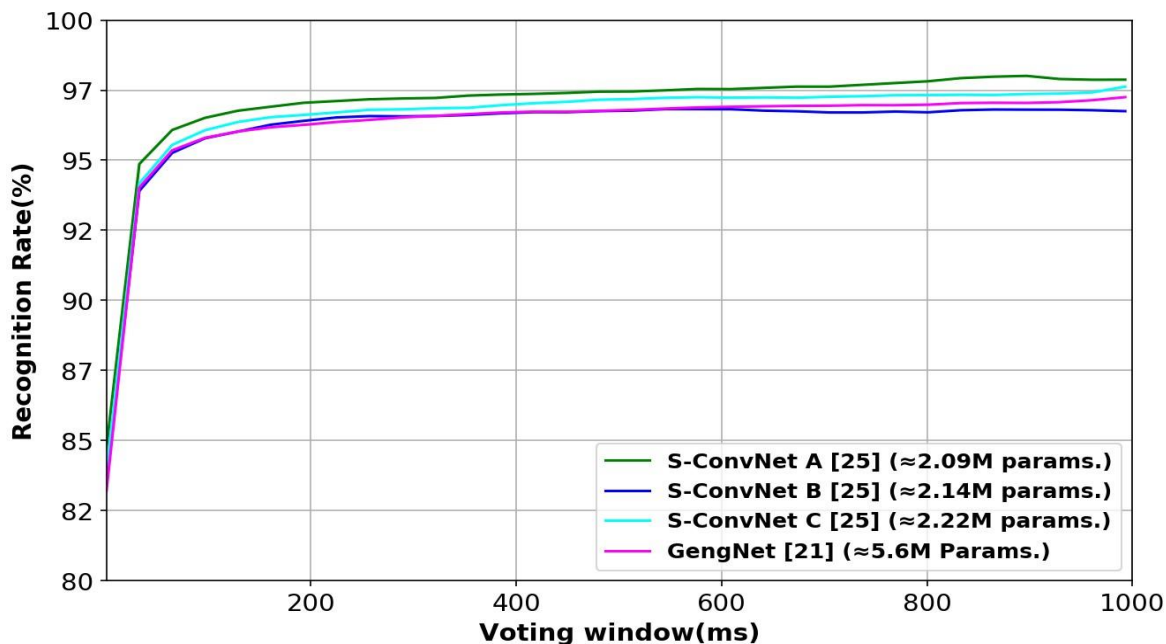
These outstanding results indicate that the proposed S-ConvNet models are highly effective for discriminative feature learning and representation in gesture recognition using instantaneous HD-sEMG images and further proved that the requirement of large-scale HD-sEMG datasets for pre-training and fine tuning by the current state-of-the-art for achieving high intra-session gesture recognition accuracy due to its deep and complex large network architecture.



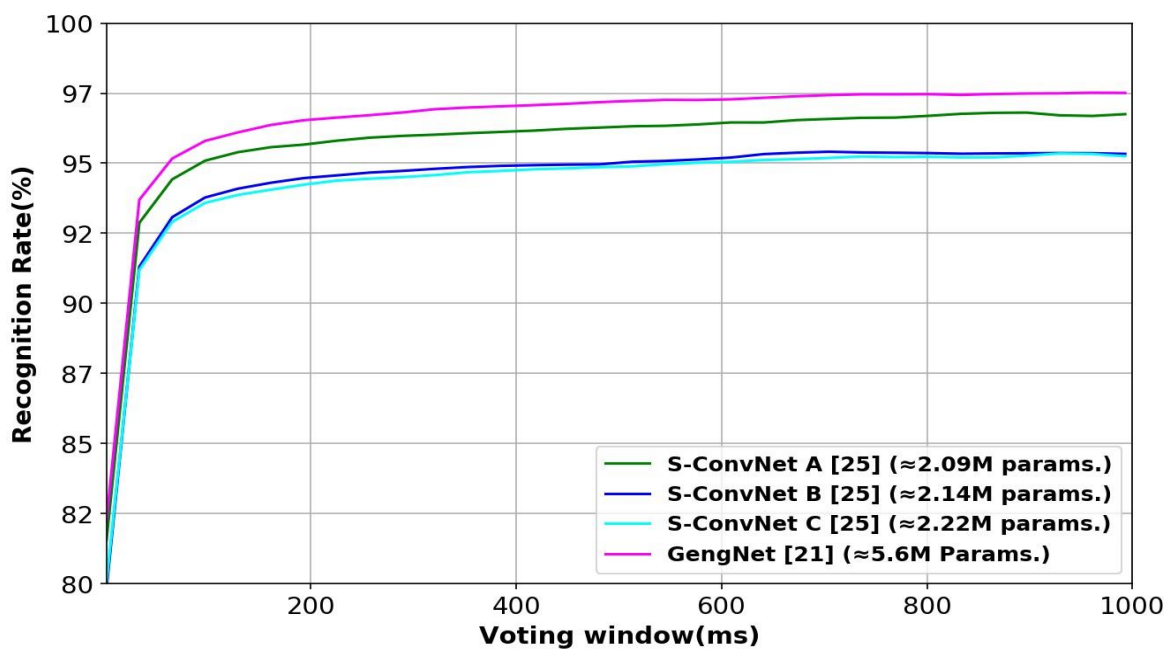
(a) CapgMyo DB-a



(b) CapgMyo DB-b (Session 1)



(c) CapgMyo DB-b (Session 2)



(d) CapgMyo DB-c

Fig. 3.6 Surface EMG gesture recognition accuracy with different voting windows using the proposed S-ConvNet models and compared with the state-of-the-art methods: (a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, (b)-(c) the gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo DB-b (Session 1) and DB-b (Session 2) respectively, and (d) the recognition accuracy of 12 hand gestures for 10 different subjects on DB-c.



### 3.4.3 Discussion on sEMG-based Gesture Recognition in Intra-Session Scenarios

The experimental results based on S-ConvNet network models demonstrate the following points:

1. the proposed S-ConvNet models trained from random-initialization can learn all the necessary invariances that requires to build a discriminant representation using only the available dataset for the target subject for gesture recognition based on instantaneous HD-sEMG images. Therefore, our discoveries will encourage community to devise shallow ConvNet architectures and train the model from the scratch (instead of pre-training) for improving the gesture recognition performance especially in the data and resource-constrained scenarios.
2. we definitely agree that, given limitless training data and unlimited computational power, deep neural networks should perform extremely well. However, our proposed approach and experimental results imply an alternative view to handle this problem: a better S-ConvNet model structure might enable similar or even better performance compared with the more complex existing models trained from large datasets by conducting an exhaustive hyperparameter search. Particularly, our S-ConvNet models are only trained with 63k instantaneous HD-sEMG images for each target subject and sets a new state-of-the-art performance for at least three (3) out of the four (4) publicly available HD-sEMG datasets, outperforming the more complex existing models trained with 720k+40K instantaneous HD-sEMG images for each of the target subject. Moreover, as the datasets grow larger, training complex deep neural networks becomes more expensive. Hence, a simple yet

efficient approach becomes increasingly significant. Despite its conceptual simplicity, our proposed methods show great potential under these settings.

3. we argue that, as aforementioned briefly, training from scratch is of critical importance at least for the following reasons. First, *Domain mismatch*– the distributions of the sEMG signals vary considerably even between recording sessions of the same subject within the same experimental set up. This problem becomes more challenging, where the learned model is used to recognize muscular activities in a new recording session. Though the fine-tuning of the pre-trained model can reduce the gap due to the deformations in a new recording session. But, what an amazing thing if we have a technique that can efficiently learn HD-sEMG images from scratch for recognizing neuromuscular activities. The proposed S-ConvNet models can be trained from scratch directly on the target subject or tasks with limited datasets and resources. Second, the fine-tuned pre-trained model restricts the structure design space for sEMG-based gesture recognition. This is very critical for the deployment of deep neural networks models to the resource limited scenarios.
4. The current state-of-the-art CNN-based gesture or neuromuscular activity recognition models require a huge memory space to store the massive parameters. Therefore, these models are usually unsuitable for low-end hand-held mobile, wearable devices and embedded electronics. Thanks to the proposed parameter-efficient S-ConvNet, our model is much smaller than the most competitive methods for instantaneous HD-sEMG image recognition. For instance, our S-ConvNet-A achieves 98.36%, 97.87%, 97.50%, and 95.80% average gesture recognition accuracy for CapgMyo DB-a, DB-b (Session 1 & 2) and DB-c datasets with a

majority voting over the recognition results of 160 frames/instantaneous sEMG images (which is equivalent to 160ms of sEMG data) with only  $\approx 2M$  parameters, which shows a greater potential for applications on low-end devices.

Drawing inspiration from the state-of-the-art intra-session sEMG-based gesture recognition accuracy achieved by the proposed S-ConvNet, the next section introduces a domain adaptation method with shallow convolutional neural network for sEMG-based gesture recognition in more challenging inter-session and inter-subject scenarios.

### **3.5 Domain Adaptation with Low-Latency Shallow Convolutional Neural Network [100]**

Experiments carried out in previous section demonstrate that, the proposed S-ConvNet models set a new state-of-the-art performance for gesture recognition in intra-session scenarios using instantaneous HD-sEMG signals. However, real-time HCIs based on sEMG-based gesture recognition in inter-session and inter-subject scenarios present a great challenge for various intrinsic and extrinsic factors, as discussed in Chapter 2, Section 2.4. In inter-session scenarios, the models are usually built by training with the data collected from the previous sessions (source domain or task) and deployed to new sessions (target domain or task) for MCI using sEMG-based gesture recognition. However, gesture recognition in inter-session scenarios based on sEMG signals is seriously hindered by the distribution shift or feature space difference between the source domain and target domain due to changes in arm posture, electrode shifts, channel variations, muscle fatigue, electrode-skin contact impedance and variations in muscle contraction force or load level [1], [11], [17], [26], [57], [63], [67], [94].

Moreover, in inter-subject scenarios, the models are usually built by training with the data collected from a group of subjects (source domain or task) and the trained model is deployed to recognize gestures from an unseen or a new subject (target domain or task). The data distribution shift caused by the above-mentioned factors in inter-session scenarios persists in inter-subject scenarios, further compounded by the added data variability arising from differences in muscle physiology between different subjects [26], [57], [63], [67], [94]. For example, when the training and test data are acquired only at different muscular contraction force or load levels, the error rate ranges from 32% to 44%. However, when the training and test data are acquired at the same muscular contraction force or load level, the error rate is reduced to a range of 8% to 19%. These findings indicate that variations in only muscular contraction force or load level can significantly impact the accuracy of an sEMG-based MCI system, with potential reductions of up to 60% [60], [105]. Therefore, the developed methods must address the distribution shift caused by these inter-session and inter-subject data variabilities for maintaining the stability of MCIs based on sEMG signals.

To address this distribution shift problem, researchers have proposed hand-crafted feature extraction and transformation-based methods with classical machine learning [15], [16], [17], [95] as discussed in Chapter 2. However, their performances are not feasible for real-time MCIs based on sEMG signals. In recent years, this distribution shift problem in sEMG-based gesture recognition in inter-session and inter-subject scenarios is addressed by deep domain adaptation (DA) methods which combines deep learning and DA [26], [57]. Domain adaptation (DA) is a form of transductive transfer learning (TL) [51] that leverages deep networks to learn transferable representations from a source domain to a target domain by embedding DA in the pipeline of deep learning.

### 3.5.1 Preliminaries

**Problem Formulation:** Given a *source domain* dataset as  $D_s = \{(x_{s_1}, y_{s_1}), \dots, (x_{s_{n_s}}, y_{s_{n_s}})\}$ , where  $x_{s_i} \in X_s$  is the data instance and  $y_{s_i} \in Y_s$  is the corresponding class label. An objective function  $f_{\theta_s}(\cdot)$  can be learned using  $D_s$  for the source task such that,  $\mathcal{T}_s = \{Y_s, f_s(\sum_i w_{s_i} X_s + b)\}$ . Similarly, we denote a *target domain* dataset as  $D_T = \{(x_{T_1}, y_{T_1}), \dots, (x_{T_{n_T}}, y_{T_{n_T}})\}$  and the task  $\mathcal{T}_T = \{Y_T, f_T(\sum_i w_{T_i} X_T + b)\}$ , where,  $x_{T_i} \in X_T$  and  $y_{T_i} \in Y_T$  are data instances and their labels respectively. In most cases, the *target domain data* is of much lower quantity compared to that of the *source domain data*, i.e.,  $0 \leq n_T \ll n_s$ . DA is a transductive TL task  $(D_s, \mathcal{T}_s, D_T, \mathcal{T}_T)$ , where the knowledge of  $D_s$  and  $\mathcal{T}_s$  is used to improve the learning of the target predictive function  $f_{\theta_T}(\cdot)$  when  $D_s \neq D_T$  and  $\mathcal{T}_s = \mathcal{T}_T$ .

In the context of sEMG-based gesture recognition problem,  $\mathcal{T}_s$  and  $\mathcal{T}_T$  refer to the same task, which involves recognizing the same set of hand gestures. However, the data distribution between  $D_s$  and  $D_T$  may deviate due to different intrinsic physiological and extrinsic environmental factors experienced in inter-session and inter-subject scenarios, as described in the previous section and Chapter 2, Section 2.4, respectively.

### 3.5.2 Baseline DA Framework and Its Limitations

Currently, Du *et al.* [26] and Ketyko *et al.* from *Nokia Bell Labs* [57] present a state-of-the-art DA solution for sEMG-based gesture recognition in inter-session and inter-subject scenarios using the CapgMyo dataset. Du *et al.* [26] propose a multi-source extension to the classical adaptive batch normalization (AdaBN) technique [37], combined with their most complex deep and large CNN architecture [21]. They employ AdaBN with the hypothesis

that the layer weights contain discriminative knowledge related to different hand gestures, while the statistics of the BatchNorm layer [57] represent discriminative knowledge from different recording sessions in inter-session or inter-subject scenarios [37]. The parameters of the pre-trained model's AdaBN [21] are updated using an unsupervised approach for adaptation in the target domain. However, a drawback of this solution arises when dealing with multiple sources (i.e., multiple subjects), as specific constraints and considerations must be imposed for each source during the pre-training phase of the model [57]. Furthermore, DA based on AdaBN achieved an inter-session gesture recognition accuracy of 63.3% for CapgMyo DB-b, and inter-subject gesture recognition accuracies of 55.3% and 35.1% for CapgMyo DB-b (session 2) and CapgMyo DB-c, respectively, by employing majority voting over the recognition results of 150 instantaneous images or frames. However, these recognition accuracies are not practical for real-time MCI applications [60]. Ketyko *et al.* [57] proposed a 2-Stage recurrent neural networks (2SRNN), where a deep stacked RNN sequence classifier was used for pre-training on the source dataset. Then, the weights of the pre-trained deep-stacked RNN classifier were frozen. At the same time, a fully connected layer without a non-linear activation function was trained in a supervised manner on the target dataset for domain adaptation. More explicitly, the deep-stacked RNN classifier was used as a feature extractor by freezing its weight in the domain adaptation stage. However, ConvNet is computationally more efficient and powerful in extracting discriminative features than RNN, even for classification tasks involving long sequences [58], [59].

### 3.5.3 Proposed Domain Adaptation (DA) Framework

The proposed domain adaptation (DA) framework for sEMG-based gesture recognition using instantaneous HD-sEMG images comprised of three key computational components: (i) S-ConvNet model development (ii) pre-training, and (iii) Domain Adaptation. A schematic diagram of the proposed DA framework for sEMG-based gesture recognition is shown in Fig. 3.7. First, we design and develop an efficient shallow convolutional neural network (S-ConvNet). The model description and its design principles are presented in Section 3.3 of this chapter.

Then, the S-ConvNet model is pre-trained from scratch on the source domain dataset. This pre-trained model encompasses a set of shared parameters denoted as  $\theta_s$ , representing the learned weights of the convolutional and fully connected layers within the S-ConvNet architecture. Additionally, it incorporates task-specific parameters, designated as  $\theta_0$  which are acquired through learning from the previously accomplished sEMG-based gesture recognition task on the source domain dataset (as illustrated in Fig. 3.7 (a)). Considering  $\theta_0$  as classifiers that operate on features parameterized by  $\theta_s$ .

The proposed domain adaptation (DA) framework comprises a feature extraction layer and a DA layer, as illustrated in Figure 3.7 (b). Regarding feature transfer to the target network for DA, this thesis introduces a feature (or weight) transfusion experiment (detailed in Chapter 4, Section 4.6.4) to explore the transferability of features within a deep convolutional neural network for sEMG-based gesture recognition. The findings suggest that feature transferability is primarily concentrated in the lower layers of the network. Drawing inspiration from these findings, the shared parameter  $\theta_s$  i.e., only the learned weights of the convolution layers acquired through pre-training the S-ConvNet model on

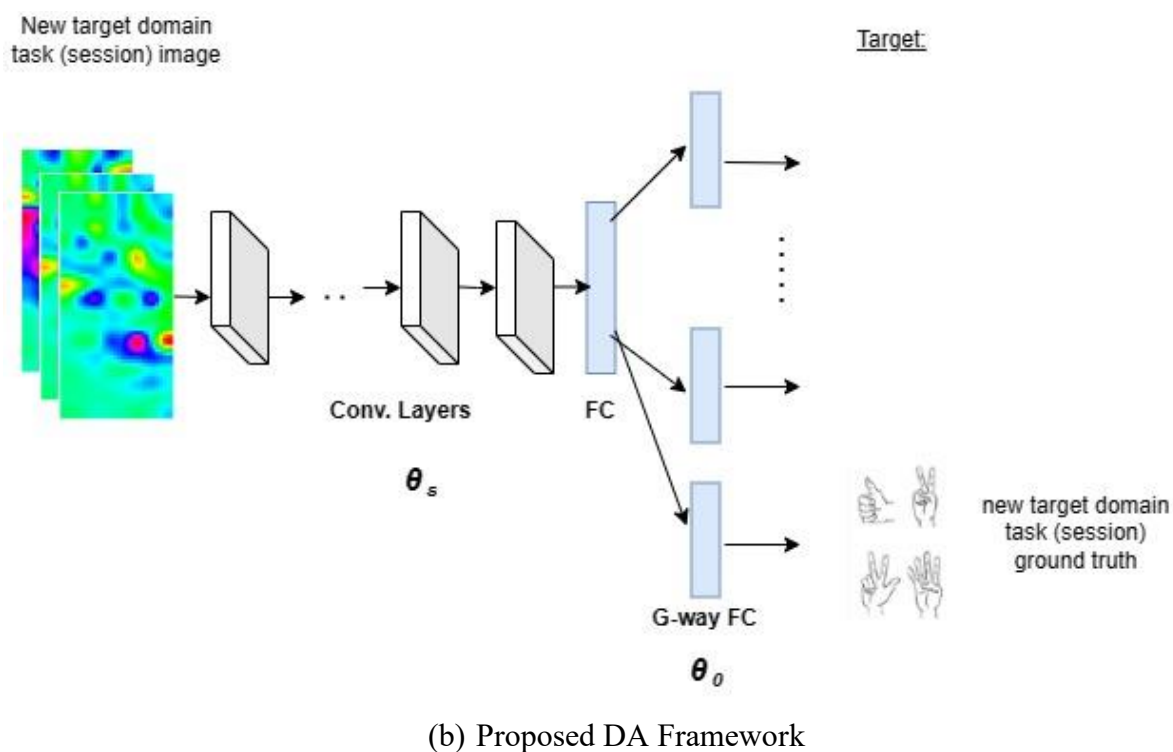
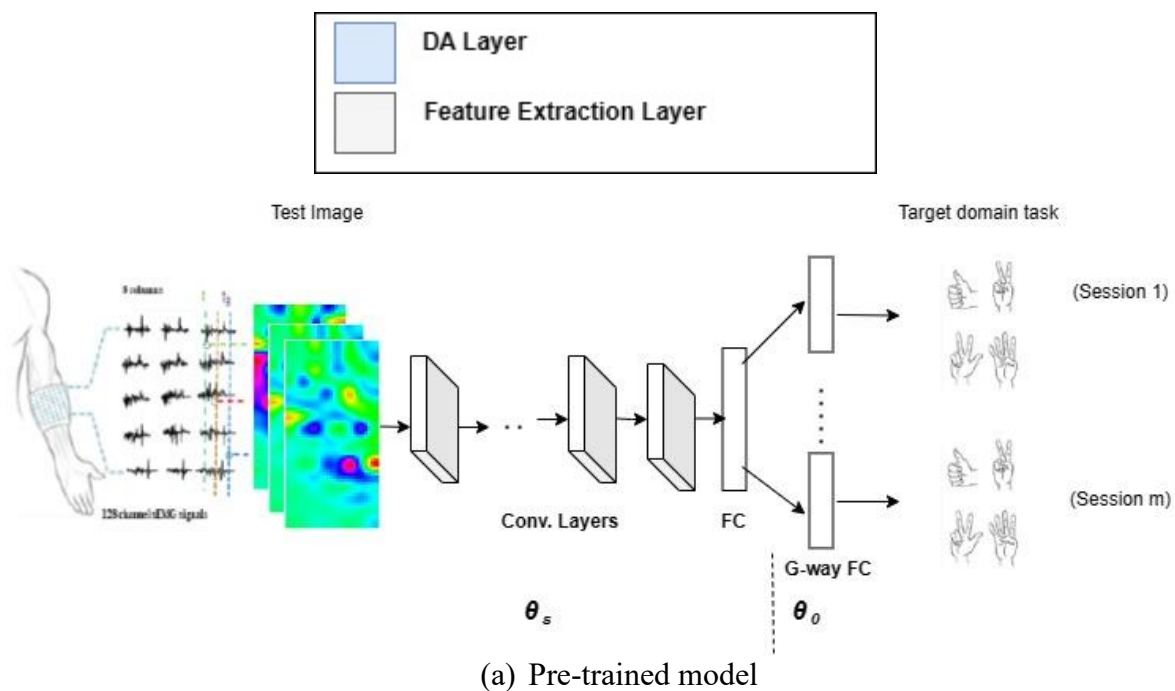


Fig. 3.7 A schematic Illustration of the proposed DA using shallow convolutional neural network (S-ConvNet). (a) Pre-trained model (b) Proposed DA Framework. SEMG images and labels used for DA are shown.



the source domain dataset, is employed as a feature extractor. This set of parameters is then transferred to a target network, enabling the target network to directly leverage the knowledge learned from the source domain. Freezing the weights of the transferred feature extractor ensures the retention of source knowledge in DA step.

Afterward, the shared parameter  $\theta_s$  of the fully connected (FC) layer, as well as the task specific parameters  $\theta_0$  of the pre-trained S-ConvNet undergo fine-tuning on the target domain dataset for DA. Finally, this adapted target network model is deployed for gesture recognition in a new session or a new subject. Essentially, it is worth mentioning that we keep the structure of the target network for DA in consistent with the proposed source S-ConvNet network model.

In the next section, we present the experimental results and training strategy for sEMG-based gesture recognition in inter-session and inter-subject scenarios and compared against the state-of-the-art.

### **3.6 Experiments in Inter-Session and Inter-Subject Scenarios**

#### **3.6.1 Experimental set up**

The proposed DA approach has been evaluated on CapgMyo dataset [26] for studying and quantifying the effects of DA on the proposed S-ConvNet network model for sEMG-based gesture recognition in inter-session and inter-subject scenarios. The CapgMyo dataset consists of three sub-datasets: CapgMyo DB-a, DB-b and DB-c. The HD-sEMG signals of a wide range of finger movements/gestures that encompass daily life activities were recorded with a sampling rate of 1000 Hz from 23 able-bodied participants whose ages ranged from 23 to 26 years. More details about the CapgMyo dataset are described in

Chapter 2, Section 2.5. CapgMyo DB-b was recorded in two distinct recording sessions with an interval of more than seven (7) days. The placement of the electrodes and/or rotations was varied for each recording session. Therefore, CapgMyo DB-b is used for the performance evaluation of the proposed DA approach with S-ConvNet. Whereas CapgMyo DB-b and DB-c were used for the performance evaluation in inter-subject scenarios.

The acquired HD-sEMG signals have been preprocessed. Consequently, the instantaneous HD-sEMG images have been generated based on the methods discussed in Section 3.2. The S-ConvNet model was pre-trained on the source dataset based on the same training strategy illustrated throughout Section 3.3 and 3.4 of this Chapter. We have also implemented the state-of-the-art network architecture [21], [26] and apply the DA to ensure a fair comparison with our proposed DA with S-ConvNet Model. However, we have adopted the same network initialization method, optimization algorithm, and training paradigm as illustrated in [21], [26]. For performance evaluation, the ConvNet model proposed in [21], [26], is considered as baseline model because this model was also employed in [23], [24], [61], and achieved the current state-of-the-art results on various sEMG-based gesture recognition datasets and tasks.

In the next subsections, the performance of the proposed DA approach is evaluated for sEMG-based gesture recognition in inter-session and inter-subject scenarios and compared against the baseline methods.

### **3.6.2 sEMG-Based Gesture Recognition in Inter-Session Scenarios**

This section presents the performance of the proposed DA with S-ConvNet and compared against the state-of-the-art for sEMG-based gesture recognition in inter-session scenarios.

In this experiments, CapgMyo DB-b is used. More explicitly, the S-ConvNet and the

compared models were pre-trained using the data recorded from the first session of CapgMyo DB-b, and the performance of the DA approach was evaluated using the second recording session of CapgMyo DB-b. DA was applied in an unsupervised manner by updating only the statistical parameters of AdaBN [37] in the target domain [26]. However, the reported very low gesture recognition accuracy as mentioned in section 3.5.2 for both inter-session and inter-subject scenarios are not enough for a usable MCI system (defined as <10% error [60]). On the other hand, DA with 2SRNN was applied in a supervised manner and reported state-of-the-art results in CapgMyo dataset [57]. Therefore, complying with the current state-of-the-art [57], the DA with S-ConvNet as well as the compared methods is applied in a supervised setting. However, DA in the context of sEMG-based gesture recognition, the main issue is *how does the developed model perform when only a small amount of data is available in the target domain for DA?* This issue must be addressed because DA often needs to be carried out under conditions of limited data availability for MCI applications based on sEMG signals, as they are often intended for amputees, elderly peoples and patients. To address this issue, we limit the available training data into five subsets: T1, T2, T3, T4 and T5, representing 20%, 40%, 60%, 80% and 100% of the total five (5) trials used for DA. The odd-numbered trials performed by the target subject in the target domain in CapgMyo dataset were used for DA, while the remaining 5 even-numbered of trials performed by the same target subject were reserved for validation. To ensure a fair comparison and align with the state-of-the-art, DA is conducted for a duration of 100 epochs. Table 3.4 presents the inter-session average gesture recognition accuracies (%) of 8 hand gestures for 10 different subjects respectively for CapgMyo DB-b and compared with the state-of-the-art methods.

Table 3.4 Inter-session gesture recognition accuracies on CapgMyo DB-b. The average recognition accuracies (%) of 8 hand gestures for 10 different subjects respectively. The numbers are the majority voted results using 150 ms window (i.e., 150 frames).

Methods	Number of available trials for DA				
	T1	T2	T3	T4	T5
Du et. al. [21][26]	68.06	81.55	86.28	88.55	88.51
2SRNN [57]	–	–	–	–	83.80
DA with S-ConvNet (Proposed)	<b>76.58</b>	<b>90.71</b>	<b>93.51</b>	<b>94.87</b>	<b>95.76</b>

Our proposed DA methods with S-ConvNet enhance inter-session gesture recognition accuracy, achieving an 11.96% improvement compared to 2SRNN [57] and a 7.25% improvement compared to GengNet [21][26] when all available 5 trials are used for DA (as shown in Table 3.4, column-T5). We also compared our proposed DA methods with S-ConvNet against the state-of-the-art GengNet [21][26], when the limited data were available for DA. The proposed DA with S-ConvNet shows even more significant improvement over the state-of-the-art when fewer trials are made available for DA, as seen in Table 3.4, Column- T1, T2, T3, and T4, respectively. For example, the proposed DA with S-ConvNet achieved an 8.52% improvement over GengNet [21][26] when only 20% of the data (i.e., 1 trial) is available for adaptation (Table 3.4, Column- T1).

### 3.6.3 sEMG-Based Gesture Recognition in Inter-Subject Scenarios

This section presents the performance of the proposed DA with S-ConvNet and compared against the state-of-the-art DA methods for sEMG-based gesture recognition in inter-subject scenarios. In this experiment, the second recording sessions of CapgMyo DB-b and

CapgMyo DB-c is used. A leave-one-subject-out cross-validation (LOSOV) is carried out, where each subject was used in succession as the target test subject, while the S-ConvNet is pre-trained using the data from the remaining subjects. Then, this pre-trained S-ConvNet model is deployed, and apply DA on the data from the odd numbers of trials of the target test subject. Finally, the adapted model was evaluated and tested using the data from the even number of trials of the target test subject. We limited the available training data to 20%, 40%, 60%, 80%, and 100% of the total 5 trials used for DA and these trials were categorized as T1, T2, T3, T4, and T5, respectively. The remaining 5 trials were kept for validation. Table 3.5 presents the average recognition accuracies (%) of 8 and 12 hand gestures for CapgMyo DB-b and DB-c for 10 different subjects, respectively.

As can be seen from Table 3.5, the proposed DA methods with S-ConvNet outperformed the state-of-the-art methods in the inter-subject scenario on both the CapgMyo DB-b and CapgMyo DB-c datasets, respectively. The proposed DA methods with S-ConvNet demonstrates an improvement of 5.33% and 7.23% compared to 2SRNN [57], and 3.78% and 4.54% compared to GengNet [21][26] on CapgMyo DB-b and CapgMyo DB-c datasets, respectively when all available 5 trials are used for DA (as shown in Table 3.5, column-T5 for both CapgMyo DB-b and CapgMyo DB-c).

Similar to the inter-session scenario, we also compared the proposed DA methods with S-ConvNet in a data starved conditions and compared the performance against the state-of-the-art GengNet [21], [26] in inter-subject scenarios. The proposed DA methods with S-ConvNet exhibits improvement over the state-of-the-art on CapgMyo DB-b and CapgMyo DB-c datasets when a limited number of trials are available for adaptation, as observed in Table 3.5, specifically in Columns T1, T2, T3, and T4, respectively. For example, when

only 20% of the data (i.e., 1 trial) was available for adaptation, the proposed DA methods with S-ConvNet achieved a 4.71% improvement over GengNet [21], [26] on CapgMyo DB-b (Table 3.5, Column- T1).

Table 3.5 Inter-subject gesture recognition accuracies. The average recognition accuracies (%) of 8 hand gestures for CapgMyo DB-b and 12 hand gestures for CapgMyo DB-c for 10 different subjects respectively. The numbers are the majority voted results using 150 ms window (i.e., 150 frames).

Methods	CapgMyo DB-b				
	Number of trials available for DA				
	T1	T2	T3	T4	T5
Du et. al. [21][26]	71.30	86.38	88.58	90.44	91.45
2SRNN [57]	-	-	-	-	89.90
DA with S-ConvNet (Proposed)	<b>76.01</b>	<b>89.58</b>	<b>92.80</b>	<b>93.98</b>	<b>95.23</b>
	CapgMyo DB-c				
Du et. al. [21][26]	<b>57.40</b>	76.30	82.45	86.15	88.09
2SRNN [57]	-	-	-	-	85.40
DA with S-ConvNet (Proposed)	57.05	<b>79.15</b>	<b>86.89</b>	<b>90.67</b>	<b>92.63</b>

We summarise the inter-session and inter-subject improvement results by the proposed DA method with S-ConvNet in Table 3.6 over the state-of-the-art DA methods. As indicated there, the performance of the proposed DA method with S-ConvNet is superior in all cases. The improvement achieved by the proposed DA method with S-ConvNet in inter-session and inter-subject scenarios, exceeds those obtained through alternative state-of-the-art domain adaptation approaches.

Table 3.6 Inter-session and Inter-subject improvement (%) results obtained by the proposed DA with S-ConvNet.

Methods	Inter-session improvement	Inter-subject improvement	
	DB-b	DB-b	DB-c
Du et. al. [21][26]	7.25	5.33	7.23
2SRNN [57]	11.96	3.78	4.54

### 3.6.4 Discussion on sEMG-Based Gesture Recognition in Inter-Session/Inter-Subject Scenarios

The need for an adequate amount of labeled data in both the source domain (training) and the target domain (test) datasets, as well as the presence of distribution shifts between these datasets, are the two main factors that hinder the successful application of deep learning to sEMG-based gesture recognition tasks. Domain adaptation (DA) with shallow convolutional neural network (S-ConvNet) effectively addresses these challenges, involving strategic knowledge transfer, optimal model adaptation, and improved generalization. The proposed DA methods leverage S-ConvNet to learn transferable representations on the source domain and adapt them to the target domain, even with very limited data available, thus demonstrating enhanced generalization capabilities. Experiments were conducted on publicly available benchmark CapgMyo datasets both in inter-session and inter-subject scenarios and compared against the state-of-the-art methods. The proposed DA methods with S-ConvNet set a new state-of-the-art performance on CapgMyo DB-b for inter-session and CapgMyo DB-b (session 2) and DB-c for inter-subject gesture recognition based on sEMG signals.

The inter-session gesture recognition accuracy achieved a notable 95.76% on CapgMyo DB-b, demonstrating a significant improvement of approximately 11.96% and 7.25% than the current state-of-the-art [57] and [21][26], respectively. In addition, the inter-subject gesture recognition accuracy reached 95.23% and 92.63% on CapgMyo DB-b and DB-c, respectively, which is 3.78% and 4.54% higher than [57] and 5.33% and 7.23% higher than the [21],[26] respectively.

These outstanding state-of-the-art inter-session and inter-subject gesture recognition accuracy validates that the proposed DA methods with S-ConvNet is highly effective in learning discriminative and domain-invariant representations to address the distribution shift caused by inter-session and inter-subject data variability.

### **3.7 Conclusion**

The requirement of sufficient amount of labeled data, high-end computational resources and distribution shift in inter-session and inter-subject scenarios are the major factors that impede deploying deep learning for real-time sEMG-based gesture recognition tasks. We aimed to address these issues in this Chapter. We present S-ConvNet models, a simple yet efficient framework for learning instantaneous HD-sEMG images from scratch for sEMG-based gesture recognition. Without using any pre-trained models, our proposed S-ConvNet demonstrate state-of-the-art performance on three (3) out of four (4) publicly available benchmark HD-sEMG datasets, while using  $\approx 12\times$ smaller dataset and reducing learning parameters to only  $\approx 2M$  for sEMG-based gesture recognition in intra-session scenarios. In addition, to address the challenging distribution shift problem, a domain adaptation method with shallow convolutional neural network is proposed. DA with shallow convolutional neural network outperformed the most complex current state-of-the-art by a large margin



both when the data from the single or multiple trials are available for adaptation for inter-session and inter-subject gesture recognition. The state-of-the-art performance on various HD-sEMG datasets and tasks proved that the proposed methods are highly effective for learning discriminative and domain-invariant representations for instantaneous HD-sEMG image recognition, especially in the data and high-end resource constrained scenarios. The proposed methods have a great potential for deploying optimal control of MCIs based on sEMG signals.

To enable on-device inference on mobile, wearable, and edge devices for sEMG-based MCIs, the model often requires periodic updates and adaptations over-the-air from the cloud CPU/GPU servers. Communication, memory and computational overhead between these mobile wearables and edge devices is directly proportional to the number of parameters in the model. With this in mind, and aiming to propose more efficient and lightweight models with fewer parameters while addressing the challenging distribution shift problem (e.g., due to electrode shift and rotations, non-uniform muscle force or contraction etc.), but outperforming the most complex current state-of-the-art deep learning model or equivalent accuracy, the next chapter presents a lightweight All-ConvNet and transfer learning framework for discriminative and domain-invariant feature representation for improved sEMG-based gesture recognition in intra-session, inter-session and inter-subject scenarios.

## **Chapitre 4 - Surface EMG-Based Inter-Session/Inter-Subject Gesture Recognition by Leveraging Lightweight All-ConvNet and Transfer Learning**

Gesture recognition using low-resolution instantaneous high-density surface electromyography (HD-sEMG) images opens up new avenues for the development of more fluid and natural muscle-computer interfaces. However, the data variability between inter-session and inter-subject scenarios presents a great challenge. The existing approaches employed very large and complex deep ConvNet or 2SRNN-based domain adaptation methods to approximate the distribution shift caused by these inter-session and inter-subject data variability. Hence, these methods also require learning over millions of training parameters and a large pre-trained and target domain dataset in both the pre-training and adaptation stages. As a result, it makes high-end resource-bounded and computationally very expensive for deployment in real-time applications. To overcome this problem, we propose a lightweight All-ConvNet+TL model that leverages lightweight All-ConvNet and transfer learning (TL) for the enhancement of inter-session and inter-subject gesture recognition performance. The All-ConvNet+TL model consists solely of convolutional layers, a simple yet efficient framework for learning invariant and discriminative representations to address the distribution shifts caused by inter-session and inter-subject data variability. Experiments on four datasets demonstrate that our proposed methods

outperform the most complex existing approaches by a large margin and achieve state-of-the-art results on inter-session and inter-subject scenarios and perform on par or competitively on intra-session gesture recognition. These performance gaps increase even more when a tiny amount (e.g., a single trial) of data is available on the target domain for adaptation. These outstanding experimental results provide evidence that the current state-of-the-art models may be overparameterized for sEMG-based inter-session and inter-subject gesture recognition tasks.

#### 4.1 Introduction

The current state-of-the-art methods [21], [23], [24], [36], [61] for sEMG-based gesture recognition either employed very complex deep and wide CNN or an ensemble of these complex networks for improved sEMG-based gesture recognition performance. For example, Geng et al. [21] exploited a DeepFace [35] like very large and deep CNN (dubbed as GengNet), which requires learning  $>5.63M$  (million) training parameters only during fine-tuning and pre-trained on a very large-scale labeled sEMG training datasets. The complexity of this model grows linearly as the input size is increased due to the use of an unshared weight strategy [27]. Wei et al. [23] used an ensemble of eight (8) single-stream GengNet at the decomposition stage only. Hu et al. [24], used a two-stage ensemble network in which an ensemble of multiple single-stream GengNet was used for spatial feature learning, resulting in multiple sequences of 1-D feature representation. Then, these 1-D feature sequences were passed to an ensemble of LSTM networks before a SoftMax layer recognized the targeted gesture. Chen et. al. [36] employed 3D CNN for learning spatial and temporal representation of sEMG images. However, the employed 3D CNN requires learning over at least  $> 30 M$  parameters, which is impractical for real-time MCI

applications based on sEMG signals. Despite the significant performance boost achieved by these state-of-the-art models [21], [23], [24], the heavy computational and intensive memory cost hinders deploying them on resource-constrained embedded and mobile devices for real-time applications. Therefore, the demand for designing low-cost, lightweight networks is highly increasing for low-end resource-limited embedded, mobile and wearable devices.

To overcome these problems, low-latency and parameter-efficient S-ConvNet is introduced in the last Chapter. More details of the proposed low-latency and parameter-efficient S-ConvNet, along with its performance comparison to the current state-of-the-art models for sEMG-based gesture recognition in intra-session, inter-session and inter-subject scenarios are described in Chapter 3. Striving to find a simpler and more efficient lightweight network, in this chapter, a new architecture called All-ConvNet is introduced that consists solely of convolutional layers and is designed to be more efficient and less computationally intensive than the existing state-of-the-art models for sEMG-based gesture recognition. Comparing the performance of All-ConvNet to other state-of-the-art models shows that it achieves competitive or state-of-the-art performance on current benchmark HD-sEMG datasets [26], while being significantly lighter, more efficient, and faster to train and evaluate. *All-ConvNet was designed based on the finding of fact that if the sEMG image area covered by units in the topmost convolutional layer covers a portion of the image large enough to recognize its content (i.e., gesture class we want to recognize).* This leads to predictions of sEMG image classes at different positions which can then simply be averaged over the whole image. Hence, the All-ConvNet becomes robust to translations and

geometric distortions, which can be very effective in addressing the electrode shift and positioning problem in sEMG-based gesture recognition.

In addition, the sEMG-based gesture recognition problem becomes more challenging in the operational conditions or an inter-session scenario, where the trained model is used to recognize muscular activities in a new recording session because sEMG signals are highly subject-specific. Inter-session is also referred to as inter-subject when the training and test data are acquired from different subjects. The distributions of the sEMG signals vary considerably in both inter-session and inter-subject scenarios due to different physiological and environmental intrinsic and extrinsic factors, as illustrated in Chapter 2, Section 2.4.

To attenuate these distribution shifts between different sEMG recording sessions, the pre-trained models have been pre-dominantly adopted by the existing approaches [26], [31], [32], and [57] to reduce the distribution shift by fine-tuning the sEMG data recorded in the different session (target domain or task). Fine-tuning updates the parameters of the pre-trained models to train to newly recorded sEMG data. Generally, the output layer of the pre-trained models is extended with randomly initialized weights. A small learning rate is used to fine-tune all the parameters from their original values to minimize the loss on the newly recorded sEMG data. Using appropriate hyper-parameters for training, the resulting fine-tuned model often outperforms learning from a randomly initialized network [40].

Generally, this pre-training and fine-tuning process can be considered a special case of domain adaptation when the source task and the target task are the same or transfer learning when the tasks are different. However, in this Chapter, we reframed this problem as transfer learning when the sEMG data for training and inference are recorded at a different session.

Fig. 4.1 illustrates the conceptual diagram of our proposed transfer-learning methods for sEMG-based gesture recognition.

Transfer learning is typically performed by taking a standard architecture along with its pre-trained weights and then fine-tuning the target task. However, the state-of-the-art methods [21], [23], [26], and [61] for sEMG-based gesture recognition employed very large and deep pre-trained models, therefore, containing millions of parameters which are designed to be trained with large-scale labeled sEMG datasets. The requirement of high-end computing resources and large-scale *pre-trained* datasets are also bounded by large and deep network structures [25]. As far as we are aware, there has been no research for sEMG-based gesture recognition studying the effects of transfer learning on the smaller, simpler, and lightweight CNN. This line of investigation is especially crucial in the sEMG-based gesture recognition because the pre-trained model is often deployed in real-time MCI applications such as assistive technology and physical rehabilitation where fine-tuning in the target domain must be conducted in the data-starved condition because of the difficulty of acquiring data from the amputees, elderly peoples, and patients, etc. Also, the large computationally expensive models might significantly impede mobile and on-device applications, where power consumption, data memory, and computational speed are constraints. To investigate the effects of transfer learning for sEMG-based gesture recognition, our research is motivated by the following research questions- *does feature reuse takes place during fine-tuning or transfer learning? And if yes, where exactly is it in the network?*

Investigating feature reuse, we find out that some of the differences from transfer learning are due to the over-parametrization of the state-of-the-art, more complex pre-trained models

rather than sophisticated feature reuse. Additionally, we discovered that a simple, lightweight model can outperform the more complex and computationally demanding state-of-the-art network architectures. We isolate where useful feature reuse occurs and outline the implications for more efficient lightweight model exploration.

In this chapter, we perform a fine-grained study on fine-tuning and transfer learning for sEMG-based gesture recognition. Our main contributions are:

- (1) We introduce All-ConvNet+TL model, which leverages the lightweight All-ConvNet and transfer learning to address the distribution shift in inter-session and inter-subject sEMG-based gesture recognition and evaluate it against the more complex state-of-the-art network architectures. Our proposed method leveraging lightweight All-ConvNet and transfer learning outperforms the state-of-the-art methods by a large margin, both when the data from a single trial or multiple trials are available for fine-tuning/adaptation. The outstanding inter-session and inter-subject gesture recognition performance achieved by the proposed lightweight models raises the question of whether the current state-of-the-art models are overparameterized for the sEMG-based gesture recognition problem.
- (2) Using further analysis and weight transfusion experiments, where we partially reuse pre-trained weights, we identify locations where meaningful feature reuse occurs and explore hybrid approaches to transfer learning. These approaches involve using a subset of pre-trained weights and redesigning other parts of the network to make them more lightweight.
- (3) We conducted more extensive experiments. A performance evaluation on four (4) publicly available HD-sEMG datasets was performed on three different sEMG-

based gesture recognition tasks: *intra-session*, *inter-session*, and *inter-subject* scenarios. The results showed that our lightweight models outperformed the more complex state-of-the-art models on various tasks and datasets.

The rest of the chapter is structured as follows: Section 4.2 presents the proposed transfer learning framework, while Section 4.3 presents the lightweight All-ConvNet model architecture and its design principles. Section 4.4 introduces the proposed transfer learning design methodology by leveraging lightweight All-ConvNet (All-ConvNet+TL). Section 4.5 describes the experimental framework, and Section 4.6 demonstrates the state-of-the-art results for inter-session and inter-subject gesture recognition and very competitive results for intra-session gesture recognition, obtained from experiments conducted on CapgMyo and its four (4) sub-datasets. Section 4.7 highlights the state-of-the-art performance achieved by the proposed All-ConvNet+TL and discusses some important findings. Finally, Section 4.8 provides some conclusive remarks.

## 4.2 The Proposed Transfer Learning Framework

The proposed transfer learning framework for sEMG-based gesture recognition using instantaneous HD-sEMG images includes the following three major computational components: (i) *a lightweight model development* (ii) *pre-training*, and (iii) *fine-tuning*. A schematic diagram of the proposed transfer learning framework for sEMG-based gesture recognition is shown in Fig. 4.1. Firstly, we devised a lightweight All-ConvNet model. Secondly, the proposed lightweight All-ConvNet was pre-trained (e.g., Fig. 4.1a) using a large amount of gesture data acquired by HD-sEMG in a single session or over multiple sessions, which may also involve multiple gestures, trials, and subjects, respectively. Then, the pre-trained model was saved and deployed for subject-specific/personalized classifier



development, as sEMG-based wearable devices are usually worn by a single user while executing a target task. Typically, input-side layers that play the role of feature extraction are copied from a pre-trained network and kept frozen or fine-tuned (e.g., Fig. 4.1b and 4.1c), in contrast, a top classifier for the target task is randomly initialized and then trained at a slow learning rate. Fine-tuning often outperforms training from scratch because the pre-trained model already has a great deal of muscular activity information. Potentially, the pre-trained network could be duplicated and fine-tuned for each new target task [40].

### 4.3 Model Description – The All-Convolutional Neural Network (All-ConvNet)

The current state-of-the-art methods [21], [23], [26], and [61] for sEMG-based gesture recognition use a large, deep ConvNet architecture similar to the one used in DeepFace [35]. This architecture is designed to be pre-trained on a large-scale labeled HD-sEMG training dataset and requires learning >5.63 million (M) parameters only during fine-tuning. As a result, this large-scale pre-trained model becomes a high-end resource-bounded and computationally very expensive to be practical for real-world MCI applications. Moreover, in their pre-trained ConvNet includes two locally connected (LCN) and three fully connected layers among the other convolutions and a  $G$ -way fully connected layer. However, the LCN layers used an unshared weight scheme [45] that makes their pre-trained ConvNet even computationally more demanding and very difficult to scale on the target domain task. The LCN layers assign an independent filter weight,  $\theta_p$  to each of the local receptive field of a feature map i.e.,  $f_p = I_p^T \theta_p$ ,<sup>22</sup> while convolution (or CNN) layers adopt a

---

<sup>22</sup>Given an input sEMG image  $I$ , LCN requires each filter is conducted on a patch vector  $I_p$ , where  $p$  stands for position of the patch in the input image.

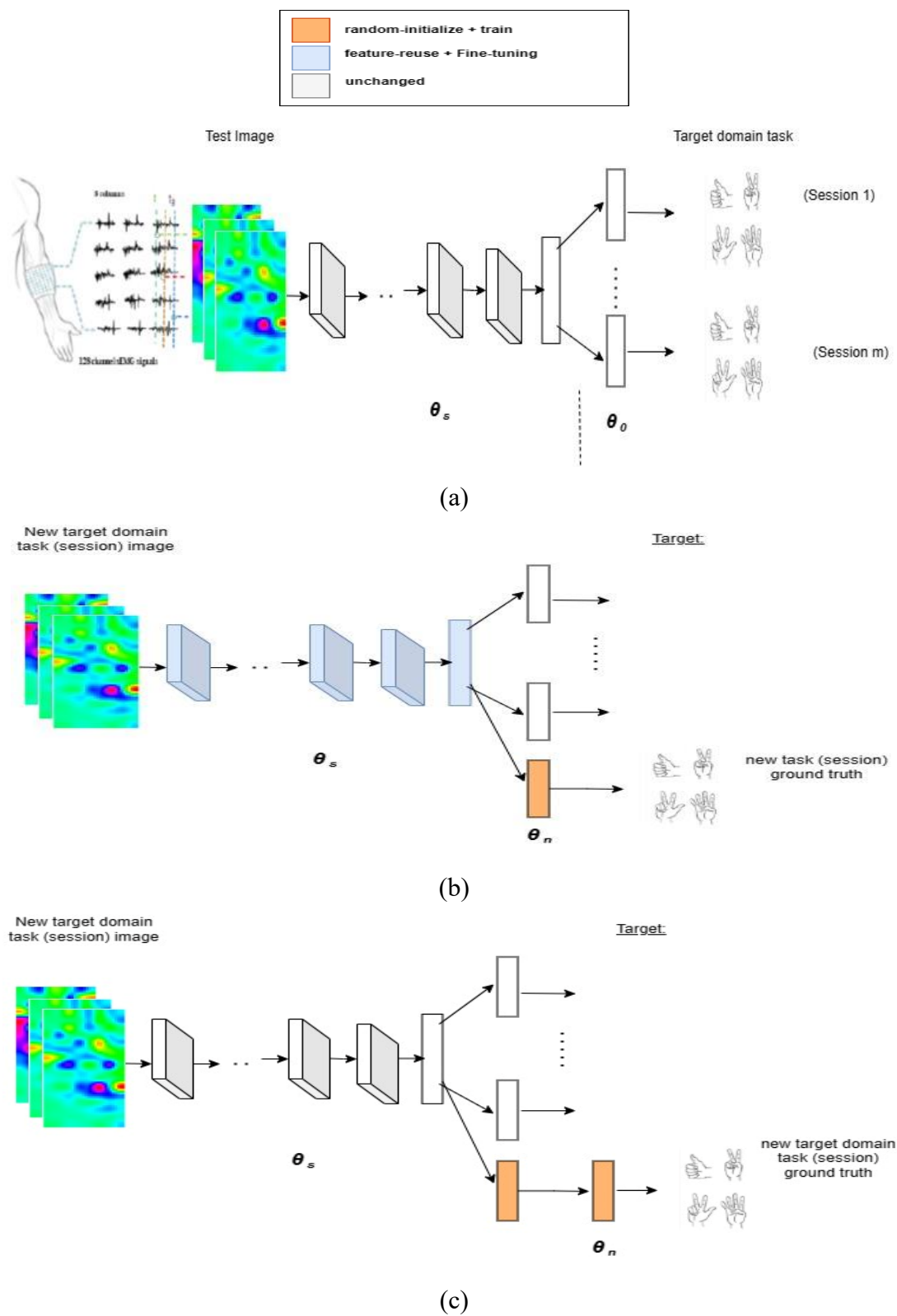


Fig. 4.1 A general conceptual diagram of the transfer learning method (a) Pre-trained model (b) Fine-tuned model and (c) Feature extraction process. sEMG images and labels used for adaptation are shown.

filter weight sharing strategy i.e.,  $f_p = I_p^T \theta$  [45]. Due to this unshared weight strategies of LCN, the number of learning parameters increases considerably from  $m$  to  $m \times k$ , where  $m \gg k$ , where  $m$  is the number of patches and  $k$  is the number of kernels. For example, the learning parameters of [21], [26] increase from  $\approx 5.63\text{M}$  to  $\approx 11\text{M}$  with a small enhancement of input HD-sEMG image size from  $16 \times 8$  to  $16 \times 16$  due to the use of this unshared weight scheme [27]. Hence, a very large-scale labeled training dataset is required for learning these growing numbers of training parameters [35]. However, the LCN can be beneficial in the application domains where the feature's precise location is dependent on the class labels. Considering the above-mentioned fact, we investigate the following research questions– (i) Do we expect the devised networks model to produce a location/translation invariant feature representation? or (ii) Do we need a location-dependent feature representation?

Following our findings and building on other recent works that aim to find a simple network architecture, we proposed a lightweight All-ConvNet. This new architecture consists solely of convolutional layers. This simple yet effective framework could learn neuromuscular activity from scratch and yield competitive or even state-of-the-art performance using a  $\approx 12 \times$  smaller dataset while reducing the learning parameters from  $\approx 5.63\text{M}$  to only  $\approx 460\text{k}$  than the more complex state-of-the-art for sEMG-based gesture recognition.

We propose a lightweight All-ConvNet, to the best of our knowledge, this is the first All-ConvNet framework to date for instantaneous HD-sEMG recognition. The All-ConvNet architectural design is adopted based on the following principles and observations:

- (i) We hypothesized that different hand gestures produce distinct spatial intensity distributions that remain consistent across multiple trials of the same gesture and distinguishable among different gestures. However, we observed that the spatial intensity distributions for the same gesture are not locally invariant, and the precise feature's location are independent of the class labels. Fig. 4.2 demonstrates a sequence of HD-sEMG images derived from the same class, which demonstrates that the distributions are independent of the class labels. CNN alone has a remarkable capability to exploit locally translational invariance features by utilizing local connectivity and weight-sharing strategies [45]. On the other hand, the LCN layer fails to model the relations of parameters in different locations. Hence, the LCN layers are ablated in designing our All-ConvNet models as the location of the features is not dependent on the class labels.

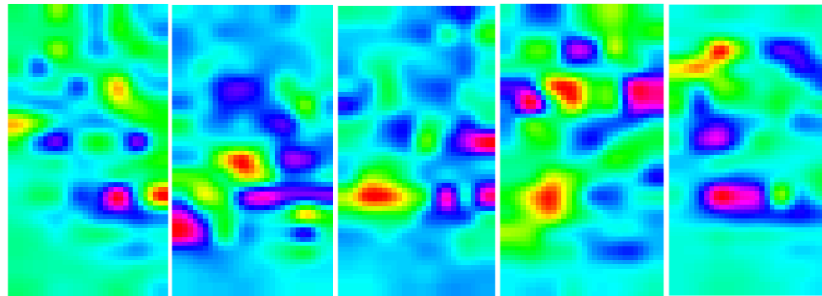


Fig. 4.2 HD-sEMGs derived from the same muscular activity class which demonstrates that the distributions are independent to the class labels.

- (ii) Inspired by previous work [46], we leverage the fact that if the part of the instantaneous HD-sEMG image is covered by the units in the topmost convolution layers could be large enough to recognize its content (i.e., the gesture class, we want to recognize). Consequently, the fully connected layers can also be replaced by

simple 1-by-1 convolutions. This allows us to predict HD-sEMG image classes at different positions, and we can then average these predictions across the entire image. Hence, the proposed All-ConvNet can be very effective in addressing the electrode shift and positioning problem for sEMG-based gesture recognition, where the entire sEMG data stream for a particular gesture may not necessarily be required for recognition. Lin *et al.* [47], initially introduced this approach, which acts as an additional regularization technique due to the significantly fewer parameters of a 1-by-1 convolution in comparison to a fully connected and LCN layers. Overall, our architecture is thus reduced to consist only of convolutional layers with ELU nonlinearities [48], [63] and a global average pooling (GAP) + SoftMax layer to produce predictions over the entire instantaneous HD-sEMG image. A conceptual diagram of our proposed pre-trained All-ConvNet is shown in Fig. 4.1(a). Table 4.1 describes our proposed All-ConvNet architecture. The feature maps learned by the proposed All-ConvNet are presented in Fig. 4.3.

Table 4.1 The All-Convnet Network Model for Neuromuscular Activity Recognition.

<b>All-ConvNet</b>
Input 16×16 Gray-level Image
3 × 3 Conv.64 ELU
3 × 3 Conv.64 ELU
3 × 3 Conv. 64 ELU with stride $r = 2$
3× 3 Conv. 128 ELU
3× 3 Conv. 128 ELU
3× 3 Conv. 128 ELU with stride $r = 2$
1×1 Conv. 128 ELU
1×1 Conv. 8 ELU
global averaging over 4×4 spatial dimensions
G-way SoftMax

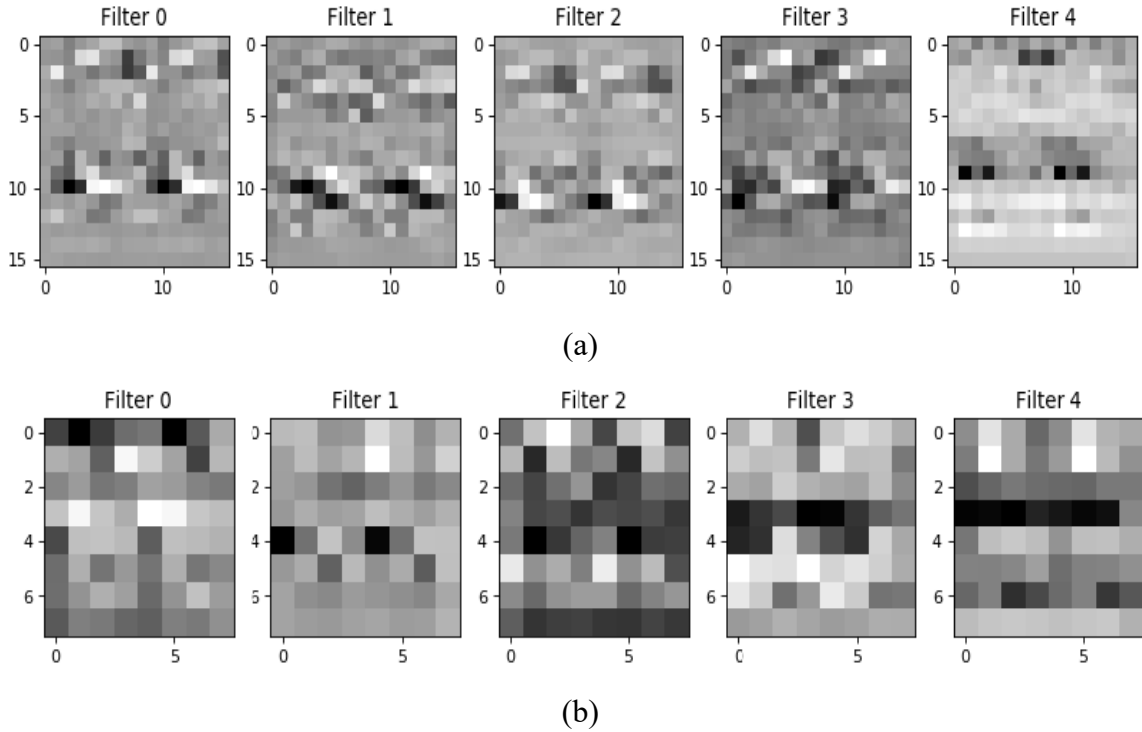


Fig. 4.3 A schematic illustration of feature maps obtained by All-ConvNet before and after dimensionality reduction. (a) Feature maps and b) Feature maps after dimensionality reduction.

We train our proposed All-ConvNet for a multi-class sEMG-based gesture recognition task, which involves recognizing a specific muscular activity class using an instantaneous HD-sEMG image. As described in Table 4.1, in the proposed All-ConvNet network, we consider using 1-by-1 convolution at the top to produce 8 or 12 outputs (depending on the number of distinct movements performed). These outputs were then averaged across all positions and fed into a  $G$ -way SoftMax layer (where  $G$  is the number of distinct hand gesture classes) which produces a distribution over the class labels. In order to estimate the class probabilities, we use the SoftMax function  $\sigma(\cdot)$  with  $\hat{y}^{(j)}$  representing the  $j$ th element of the  $G$  dimensional output vector of the layer preceding the SoftMax layer, defined as below:

$$\sigma(\hat{y}^{(j)}) = \frac{\exp(\hat{y}^{(j)})}{\sum_G \exp(\hat{y}^{(G)})} \quad (4.1)$$

The objective of this training is to maximize the probability of the correct gesture class.

This is accomplished by minimizing the cross-entropy loss [49] for each training sample.

When  $y$  represents the true label for a given input, the loss is computed as:

$$L = -\sum_j y^{(j)} \ln(\sigma(\hat{y}^{(j)})) \quad (4.2)$$

The loss is minimized over the parameters by computing the gradient of  $L$  with respect to the parameters. These parameters are then updated using the state-of-the-art Adam (adaptive moment estimation) gradient descent-based optimization algorithm [50]. This algorithm provides fast and reliable learning convergence, unlike the stochastic gradient descent (SGD) optimization algorithm used in state-of-the-art pre-trained networks for gesture recognition using instantaneous HD-sEMG image recognition.

Once the network has been trained, an instantaneous HD-sEMG image is recognized as in the gesture class  $C$  by simply propagating the input image forward and computing:

$$C = \operatorname{argmax}_j(\hat{y}^{(j)}) \quad (4.3)$$

#### 4.4 Transfer Learning by Leveraging Lightweight All-ConvNet (All-ConvNet+TL)

In this section, we introduce some notations and definitions used in our transfer learning framework as in [51]. We denote the source domain data as  $D_S = \{(x_{S_1}, y_{S_1}), \dots, (x_{S_{n_S}}, y_{S_{n_S}})\}$ , where  $x_{S_i} \in X_S$  is the data instance and  $y_{S_i} \in Y_S$  is the corresponding class label. In our sEMG-based gesture recognition example,  $D_S$  can be a set of sEMG data of different gestures and their corresponding gesture class labels acquired by a single or multiple participants in a designated session. An objective function  $f_S(\cdot)$  can be

learned using  $D_s$  for the source task such that,  $\mathcal{T}_s = \{Y_s, f_s(\sum_i w_{s_i} X_s + b)\}$ . Similarly, we denote the target domain data as  $D_T = \{(x_{T_1}, y_{T_1}), \dots, (x_{T_{n_T}}, y_{T_{n_T}})\}$  and  $\mathcal{T}_T = \{Y_T, f_T(\sum_i w_{T_i} X_T + b)\}$ , where,  $x_{T_i} \in X_T$  and  $y_{T_i} \in Y_T$  are the sEMG data of different gestures and their corresponding class labels respectively acquired by a distinct subject/participant at a different session than  $D_s$ . In most cases, the target domain data for a distinct participant acquired at another session is much lower quantities than that of a source domain data, *i.e.*,  $0 \leq n_T \ll n_s$ .

Now we define our proposed transfer learning problem as follows– Given a source domain  $D_s$  and a learning task  $\mathcal{T}_s$  as well as a target domain  $D_T$  and learning task  $\mathcal{T}_T$ , the transfer learning aims to help improve the learning of the target predictive function  $f_T(\cdot)$  in  $D_T$  using the knowledge in  $D_s$  and  $\mathcal{T}_s$ , where,  $D_s \neq D_T$ , and  $\mathcal{T}_s = \mathcal{T}_T$ . In our sEMG-based gesture recognition problem, the source and target task are the same. However, the data distribution between the source and the target domain might be different *i.e.*,  $D_s \neq D_T$  due to intrinsic and extrinsic factors described in Chapter 2, Section 2.4.

To mitigate these distribution shifts on the sEMG-based gesture recognition problem, we apply the transfer learning to our proposed lightweight All-ConvNet and termed it as All-ConvNet+TL. In our setting, All-ConvNet+TL has a set of shared parameters  $\theta_s$  (e.g., all the convolutional layers in All-ConvNet) and task-specific parameters for previously learned gesture recognition tasks  $\theta_0$  (e.g., the output layer of All-ConvNet for gesture recognition and its corresponding weights), and the task-specific parameters are randomly initialized for new target tasks  $\theta_n$  (e.g., gesture recognition in a new session). Considering  $\theta_0$  and  $\theta_n$  as classifiers that operate on features parameterized by  $\theta_s$ . Drawing motivation



from [40], [65-66], in this work, we adopt the following approaches to learning  $\theta_n$  while taking advantage of previously learned  $\theta_s$ , which is illustrated in Fig. 4.1:

- (i) ***Fine-tuning*** – involves optimizing  $\theta_s$  and  $\theta_n$  for the new target task, while keeping  $\theta_0$  fixed (as shown in Fig.4.1b). To prevent large drift in  $\theta_s$ , a low learning rate is usually used. It is possible to duplicate the original network and fine-tune it for each new target task to create a set of specialized networks.
- (ii) ***Feature Extraction*** –  $\theta_s$  and  $\theta_0$  remain fixed and unchanged, while the outputs of one or more layers are used as features for the new target task in training  $\theta_n$  (as shown in Fig. 4.1c).

The most popular methodology for transfer learning is to duplicate the pre-trained network (i.e., initialize from pre-trained weights) and fine-tune (train) the entire network for each new target task [62]. However, fine-tuning degrades performance on previously learned tasks from the source dataset because the shared parameters change without receiving new guidance for the source-task-specific prediction parameters. In addition, duplicating and fine-tuning all the parameters of a *pre-trained model* may also require a substantial amount of target task dataset. On the other hand, feature extraction usually underperforms on the target dataset because the shared parameters often fail to effectively capture some discriminative information that is crucial for the target task. To address this problem and find out a good trade-off between fine-tuning and feature extraction, we focus on answering the following research questions – *Does feature reuse take place during fine-tuning or transfer learning? And if yes, where exactly is it in the network?* We first conducted a preliminary weight (or feature) transfusion experiment, where we partially reused pre-trained weights to determine and isolate the locations where meaningful feature reuse

occurs. We perform this via a weight transfusion experiment by transferring a contiguous set of some of the pre-trained weights, randomly initializing the rest of the network, and training on the target task. We have found out that meaningful feature reuse is restricted to the lowest few layers of the network and is supported by gesture recognition accuracy and convergence speed (see in Section 4.6.4 for details). Following the results of these weight (or feature) transfusion experiments, the part of the  $\theta_s$  (i.e., the first three convolutional layers of All-ConvNet) were frozen and used as a feature extractor and only  $\theta_t$  in the top convolutional layers were fine-tuned. Hence, the proposed network model allows the target task to leverage complex features learned from the source dataset and make these features more discriminative for the target task by fine-tuning the top convolutional layers. These transfusion results suggest we propose hybrid and more flexible approaches to transfer learning (see Section 4.6.5).

#### 4.5 Experimental Setup

We evaluated our proposed approach on CapgMyo dataset [26] for studying and quantifying the effects of transfer learning on the smaller, simpler, and lightweight CNN. More details of the CapgMyo dataset are described in Chapter 2, Section 2.5. All three sub-databases of CapgMyo DB-a, DB-b, and DB-c were used for *intra-session* performance evaluation. *Inter-session* recognition of hand gestures based on sEMG typically suffers from electrode shift and positioning. Therefore, DB-b was used for *inter-session* performance evaluation. Finally, both DB-b Session 2 and DB-c were used for inter-subject performance evaluation.

For CapgMyo database, first, the power-line interferences were removed from the acquired HD-sEMG signals using a 2nd order Butterworth filter with a band-stop range between 45

and 55 Hz. Then, the HD-sEMG signals were arranged in a 2-D grid according to their electrode positioning at each sampling instant. Afterward, this grid was transformed into an instantaneous sEMG image by linearly converting the values of sEMG signals from  $mV$  to color intensity as  $[-2.5mV, 2.5mV]$  to  $[0, 255]$ . As a result, instantaneous grayscale sEMG images with a size of  $16 \times 8$  matrices were obtained. To facilitate GAP, we enhance the input HD-sEMG image size from  $16 \times 8$  to  $16 \times 16$  using horizontal mirroring. Unlike [21], [26] this enhancement does not increase the learning parameters in the proposed All-ConvNet.

For pre-training our proposed original model All-ConvNet, the following configurations is adopted, the connection weights for All-ConvNet network architecture were randomly initialized using Xavier initialization scheme [52], [53] and the network was trained using Adam optimization algorithm [50]. The momentum decay and scaling decay were initialized to 0.9 and 0.999, respectively. In contrast to SGD employed in [21], [23], and [26], Adam is an adaptive learning rate algorithm, therefore it requires less tuning of the learning rate hyperparameter. For all our experiments, the learning rate of 0.001 was initialized, and smaller batches of 256 randomly chosen samples from the training dataset were fed to the network during consecutive learning iterations. We set a maximum of 100 epochs for training our All-ConvNet model. However, to prevent overfitting, we applied early stopping [54], which interrupts the training process if no improvements in validation loss are observed for 5 consecutive epochs. BN [55] was applied after the input and before each non-linearity. To further regularize the network, Dropout [56] was applied to all layers with a probability of 25%. The All-ConvNet model was trained on a workstation with an Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz processor, 32 GB RAM, and an NVIDIA

RTX 2080 Ti GPU. Each epoch was completed in approximately 6s for a test on intra-session gesture recognition. We have also implemented the state-of-the-art network architecture [21] for a fair comparison with our proposed lightweight sEMG-based gesture recognition algorithm. However, we have adopted the same network initialization method, optimization algorithm, and training paradigm as illustrated in [21].

## 4.6 Experimental Results

From the viewpoint of MCI application scenarios, the sEMG-based gesture recognition can be categorized into three (3) scenarios such as intra-session, inter-session and inter-subject scenarios as described in Chapter 2, Section 2.4 and in Chapter 3, Section 3.4 respectively. However, the sEMG-based gesture recognition methods in the literature have usually been investigated in intra-session scenarios [21], [23], [24], [36] and [61]. Similar to Chapter 3, we evaluated the performance of our proposed sEMG-based gesture recognition algorithm by leveraging lightweight All-ConvNet and transfer learning in inter-session and inter-subject scenarios in addition to intra-session gesture recognition. In the following subsections, we evaluated the performance of our proposed lightweight gesture recognition algorithms. We compared them with the state-of-the-art, more complex methods in the above-mentioned three different scenarios.

### 4.6.1 Intra-Session Performance Evaluation

In this section, we evaluated the performance of sEMG-based gesture recognition in the intra-session scenario. In this scenario, usually, the data variation comes from the difference between the trials and repetitions of the hand/finger gestures performed by an individual. To mitigate this data variations or distribution time shift caused by the repetitions of the gestures in multiple trials in the same session, the state-of-the-art methods

performed pre-training their proposed CNN using half of the training data from all the participated subjects (e.g., 18 in DB-a) in the data collection process. Then, the pre-trained model was fine-tuned using the training data from the target subject for the subject-specific classifier development. The major drawback of this approach [21] is that the same training data used for fine-tuning was also seen during pre-training. However, we argued that the proposed lightweight All-ConvNet trained from scratch using random initialization has the great ability to model these distribution shifts caused by the repetitions of hand gestures across multiple trials within the same session. In this setting, we proposed designing and developing a subject-specific individualized classifier using only the sEMG data available for an individual or target subject while executing a target task without pre-training. For example, in CapgMyo DB-a and DB-b, eight (8) isotonic and isometric hand gestures were performed by an individual subject. Each gesture was also trialed and recorded 10 times with a 1000 Hz sampling rate. Thus, an individual subject generates ( $8 \times 10 \times 1000 = 80,000$ ) instantaneous sEMG images. In CapgMyo DB-c, an individual performed twelve (12) basic movements of the fingers, and hence it generates ( $12 \times 10 \times 1000 = 120,000$ ) instantaneous sEMG images. For performance evaluation of the proposed subject-specific lightweight All-ConvNet, a leave-one-trial-out cross-validation (LOTOV) was performed, in which each of the 10 trials was used in turn as the test set, and the proposed lightweight All-ConvNet was trained and validated using the remaining 9 trials. This entire paradigm of training and testing process is illustrated in Fig. 4.1a, which shows that only the trained model (without any feature reuse from the pre-trained model) is used for gesture recognition. In this chapter, we performed an extensive experiment on the CapgMyo DB-a, DB-b (session 1), DB-b (session 2) and DB-c, respectively. Table 4.2 presents the gesture

recognition results for the proposed lightweight All-ConvNet and compares them with the state-of-the-art methods.

As can be seen in Table 4.2, the proposed lightweight All-ConvNet (with around only 0.46 million learning parameters) consists of a stack of 3×3 convolutional layers with occasional subsampling by a stride of 2. It is trained from random initialization and outperformed the state-of-the-art, more complex GengNet [21], [23], [24], [26] and [61] on the CapgMyo DB-b Session 1 and Session 2 datasets, respectively, and performs comparably to the S-ConvNet presented in Chapter 3. Additionally, the lightweight All-ConvNet performs very competitively or on par with the GengNet [21] and S-ConvNet on the CapgMyo DB-a and CapgMyo DB-c datasets, respectively.

Table 4.2 The average recognition accuracies (%) of 8 hand gestures for CapgMyo DB-a and DB-b for 18 and 10 different subjects respectively and 12 gestures for 10 different subjects in DB-c. The numbers are the majority voted results using 160 ms window (i.e., 160 frames). Per-frame accuracies are shown in parenthesis.

<b>Model</b>	<b>S-ConvNet</b>	<b>W.Geng et. al [21]</b>	<b>All-ConvNet (proposed)</b>
CapgMyo DB-a	<b>98.36 (87.95)</b>	98.48 (86.92)	98.02 (86.73)
CapgMyo DB-b Session 1	<b>97.87 (83.57)</b>	97.04 (81.26)	97.52 (81.95)
CapgMyo DB-b Session 2	<b>97.05 (84.73)</b>	96.26 (83.21)	96.80 (83.36)
CapgMyo DB-c	<b>95.80 (81.63)</b>	96.36 (82.23)	95.76 (80.91)
#Learning Parameters	≈ 2.09 <i>M</i>	≈ 5.63 <i>M</i>	≈ <b>0.46 <i>M</i></b>

Fig. 4.4 (a)-(d) presents the sEMG-based instantaneous (or per-frame) gesture recognition accuracies and their statistical significance obtained through leave-one-trial-out cross-validation for ten different test trials for each of the participating subjects in CapgMyo

DB-a, DB-b, and DB-c, respectively. The highest instantaneous (or per-frame) gesture recognition accuracies were 86.73% for DB-a, 81.95% and 83.36% for DB-b (Session 1 and Session 2, respectively), and 80.91% for DB-c. Which were obtained with the proposed lightweight All-ConvNet. The high per-frame gesture recognition accuracies and low standard deviation over multiple test trials and subjects in each of the four HD-sEMG datasets mentioned above reflect the high stability of the proposed lightweight All-ConvNet.

In addition, based on a simple majority voting algorithm, we have obtained very good gesture recognition accuracies. Fig. 4.5 (a)-(d) presents gesture recognition accuracy with different voting windows using lightweight All-ConvNet. The average gesture recognition accuracy of 94.56% and 95.99% were achieved by a simple majority voting with 32 and 64 instantaneous images (or frames) for the above-mentioned four (4) HD-sEMG datasets.

The higher gesture recognition accuracies of 98.02%, 97.52%, 96.80%, and 95.76% (as shown in Table 4.2 and Fig. 4. 5) can be obtained by the proposed lightweight All-ConvNet and a simple majority voting over the recognition result of 160 frames for DB-a, DB-b (Session 1 and Session 2) and DB-c, respectively. These outstanding results confirm that the proposed lightweight All-ConvNet is highly effective for learning all the invariances for low-resolution instantaneous HD-sEMG image recognition and hence seem to be enough to address the problem of employing high-end resource-bounded fine-tuned pre-trained networks for low-resolution instantaneous HD-sEMG image recognition.

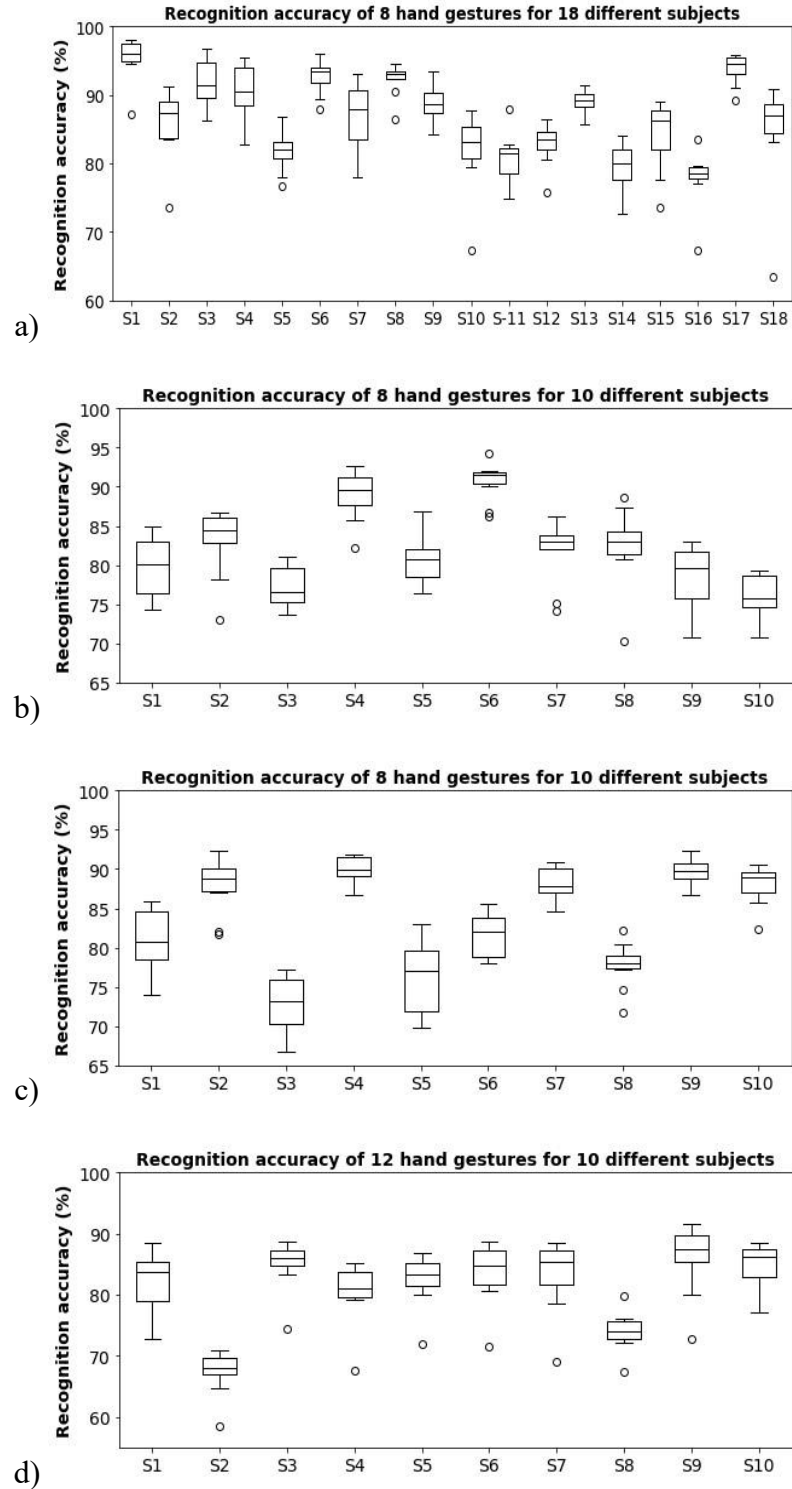
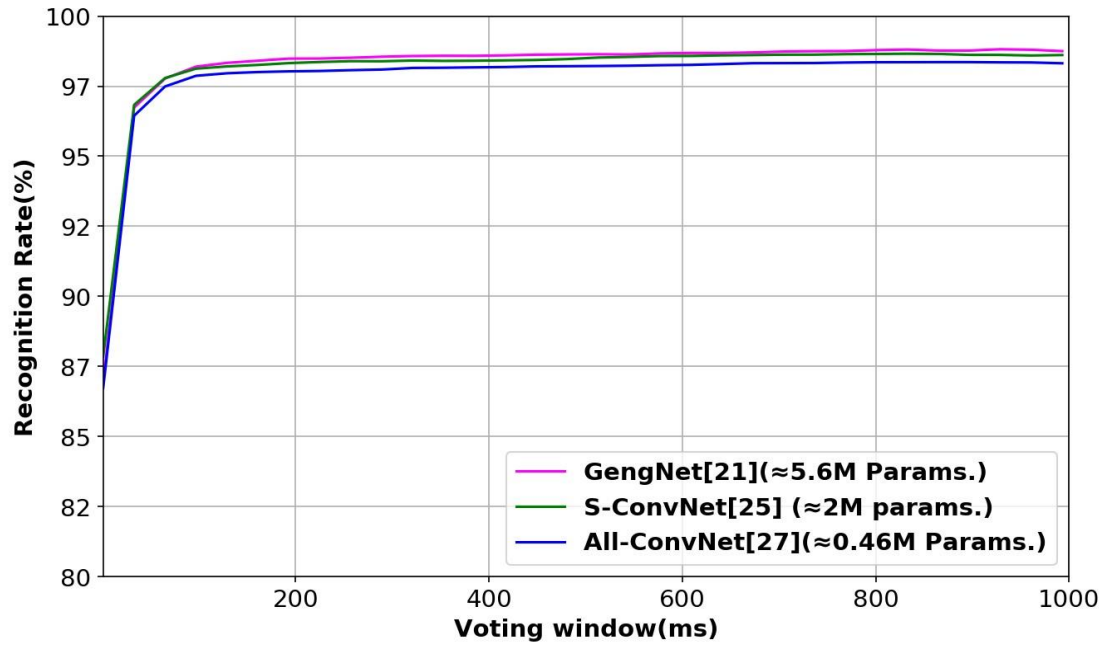
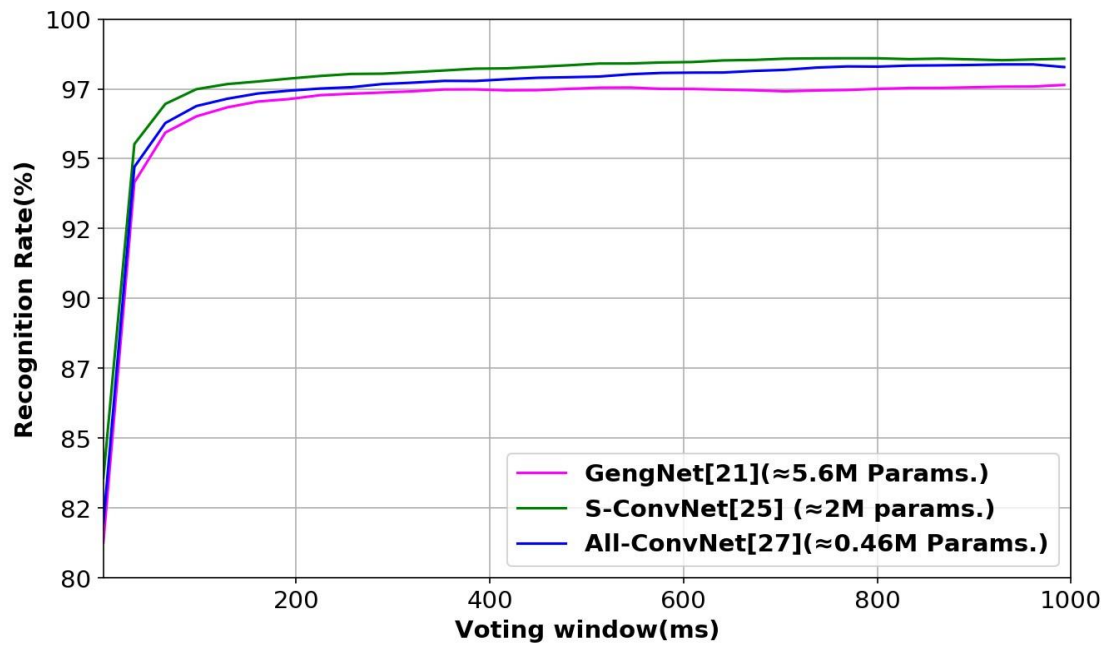


Fig. 4.4 The per-frame gesture recognition accuracy with our proposed lightweight All-ConvNet (a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, (b)-(c) The gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo DB-b (Session 1) and DB-b (Session 2) respectively (d) the gesture recognition accuracy of 12 hand gestures for 10 different subjects on CapgMyo DB-c.

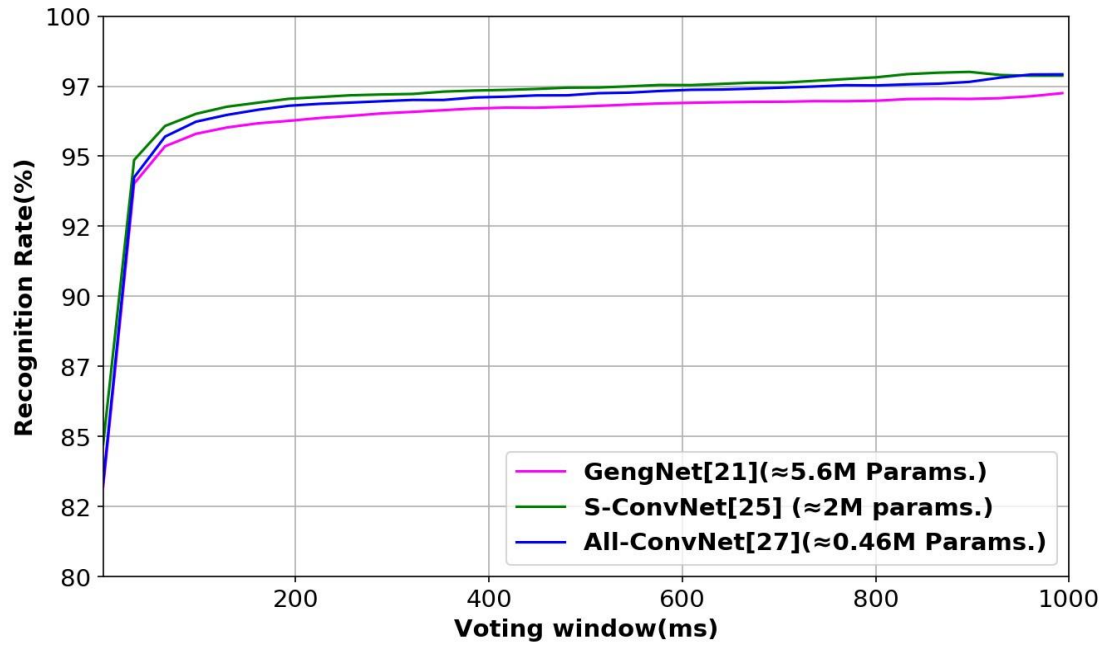




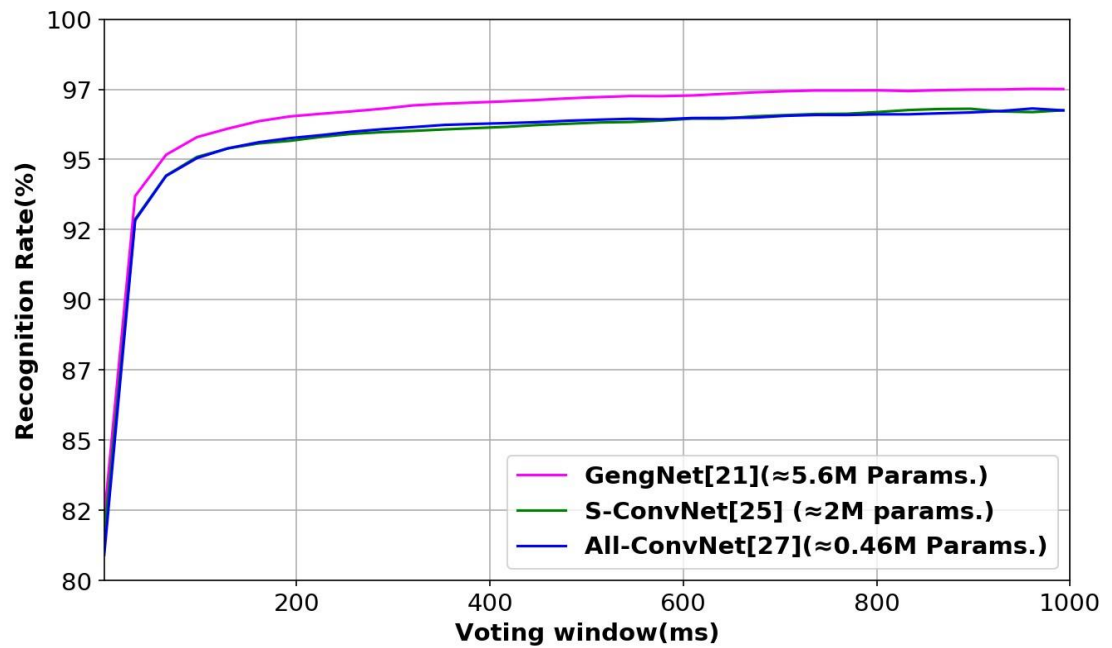
a)



b)



c)



d)

Fig. 4.5 Surface EMG gesture recognition accuracy with different voting windows using the proposed lightweight All-ConvNet and compared with the state-of-the-art methods: a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, and the gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo for b) DB-b Session 1 and c) DB-b Session 2, and d) the recognition accuracy of 12 hand gestures for 10 different subjects on DB-c.

#### 4.6.2 Inter-Session Performance Evaluation

In this section, we evaluated the performance of sEMG-based gesture recognition in the inter-session scenario. In this scenario, there is still the intra-session variability discussed in the previous section, in addition to the extent of data variability, which comes from the differences between the recording sessions. The sensor placement may have some spatial shifts and/or rotations at each recording session. These differences in sensor placement and/or rotations may cause spatial shifts in the distributions of the sEMG sensor data. To address this spatial shift problem, currently [26] and [57] provide a state-of-the-art solution in the CapgMyo dataset. Du *et al.* [26] proposed a multi-source extension to the classical adaptive batch normalization (AdaBN) technique [37] for domain adaptation, which works with CNN architecture. The drawback of this solution is that when dealing with multiple sources (i.e., multiple subjects), it is necessary to impose specific constraints and considerations for each source during the pre-training phase of that model [57]. Ketyko *et al.* [57] proposed a 2-Stage recurrent neural networks (2SRNN), where a deep stacked RNN sequence classifier was used for pre-training on the source dataset. Then, the weights of the pre-trained deep-stacked RNN classifier were frozen. At the same time, a fully connected layer without a non-linear activation function was trained in a supervised manner on the target dataset for domain adaptation. More explicitly, the deep-stacked RNN classifier was used as a feature extractor by freezing its weight in the domain adaptation stage. However, ConvNet is more powerful at extracting discriminative features than RNN, even for classification tasks of long sequences [58], [59].

In addition, it is noteworthy that the domain adaptation was conducted in unsupervised and semi-supervised settings [26]. However, very low gesture recognition accuracies were

reported in [26] in both inter-session and inter-subject scenarios. On the other hand, [57] performed domain adaptation in supervised settings and demonstrated state-of-the-art results on the CapgMyo dataset. Therefore, for a fair comparison with the state-of-the-art, we performed domain adaptation in a supervised manner in all the compared methods. Moreover, it might be an interesting question why we chose to compare the performance of our proposed lightweight All-ConvNet+TL with the CNN models, proposed in [21] and [26]. To the best of our knowledge, the base CNN models proposed in [21] and [26] were also adapted in [23], [24], and [61], respectively, and reported state-of-the-art results on various sEMG-based gesture recognition tasks and datasets.

Experiments conducted on inter-session and inter-subject settings; we have shown that our proposed lightweight All-ConvNet+TL leveraging transfer learning (illustrated in Section 4.4) outperformed these above-mentioned state-of-the-art solutions. We evaluated inter-session gesture recognition for CapgMyo DBb, in which the model was trained using data recorded from the first session and evaluated using data recorded from the second session. It is worth mentioning that without transfer learning or domain adaptation, the state-of-the-art models, as well as our proposed models achieved less than or approximately 50% average gesture recognition accuracy on CapgMyo datasets in both inter-session and inter-subject scenarios. This level of recognition accuracy is not enough for a usable system (defined as  $<10\%$  error [60]). Therefore, domain adaptation or transfer learning must be introduced to these (inter-session and inter-subject) settings for acceptable performance. However, the most significant question is how much training data is required for adaptation on the target domain to obtain a stable gesture recognition accuracy. To address this question, we limited the available training data to 20% (T1), 40% (T2), 60% (T3), 80%

(T4), and 100% (T5) of the total 5 trials used for domain adaptation (the remaining 5 trials are kept for validation). For fair comparison and complying with the state-of-the-art, we ran our domain adaptation for 100 epochs. Table 4.3 presents the inter-session average gesture recognition accuracies (%) of 8 hand gestures for 10 different subjects respectively for CapgMyo DB-b and compared with the state-of-the-art methods.

Table 4.3 Inter-session gesture recognition accuracies on CapgMyo DB-b. The average recognition accuracies (%) of 8 hand gestures for 10 different subjects respectively. The numbers are the majority voted results using 150 ms window (i.e., 150 frames).

Methods	Number of available trials for adaptation				
	T1	T2	T3	T4	T5
Du et. al. [21][26]	67.97	81.77	86.02	88.10	88.48
2SRNN [57]	-	-	-	-	83.80
All-ConvNet+TL (Proposed)	<b>75.91</b>	<b>89.61</b>	<b>92.74</b>	<b>93.46</b>	<b>94.91</b>

Our proposed lightweight All-ConvNet+TL leverages transfer learning to enhance inter-session gesture recognition, achieving an 11.11% improvement compared to 2SRNN [57] and a 6.43% improvement compared to GengNet [21], [26] when all available 5 trials are used for adaptation (as shown in Table 4.3, column-T5). We also compared our proposed lightweight All-ConvNet+TL with the state-of-the-art GengNet [21], [26] in a data-starved condition. The proposed lightweight All-ConvNet+TL shows even more significant improvement over the state-of-the-art when a limited number of trials are available for adaptation, as seen in Table 4.3, Columns- T1, T2, T3, and T4, respectively. For example, the proposed lightweight All-ConvNet+TL achieved a 7.94% improvement over GengNet

[21][26] when only 20% of the data (i.e., 1 trial) was available for adaptation (Table 4.3, Column- T1).

### 4.6.3 Inter-Subject Performance Evaluation

In this section, we evaluated the performance of sEMG-based gesture recognition in the inter-subject scenario. In this scenario, the data variability comes from the variation in muscle physiology between different subjects. In this experiment, we evaluated the inter-subject recognition of 8 gestures using the second recording session of CapgMyo DB-b and the recognition of 12 gestures using CapgMyo DB-c. We performed a leave-one-subject-out cross-validation, in which each of the subjects was used in turn as the test subject, and a lightweight All-ConvNet was pre-trained using the data of the remaining subjects. Then, this pre-trained All-ConvNet model was deployed, and adaptation was made on the data from the odd numbers of trials of the test subjects by leveraging transfer learning or domain adaptation. Finally, the adapted model was evaluated and tested using the data from the even number of trials of the test subject. We limited the available training data to 20%, 40%, 60%, 80%, and 100% of the total 5 trials used for domain adaptation (the remaining 5 trials are kept for validation). Table 4.4 presents the average recognition accuracies (%) of 8 and 12 hand gestures for CapgMyo DB-b and DB-c for 10 subjects, respectively.

As can be seen from Table 4.4, our proposed lightweight All-ConvNet+TL, by leveraging transfer learning, outperformed the state-of-the-art methods in the inter-subject scenario on both CapgMyo DB-b and CapgMyo DB-c datasets, respectively. Our proposed lightweight All-ConvNet+TL demonstrates an improvement of 5.04% and 6.17% compared to 2SRNN [57], and 3.58% and 1.85% compared to GengNet [21], [26] on CapgMyo DB-b and

CapgMyo DB-c datasets, respectively when all available 5 trials are used for adaptation (as shown in Table 4.4, column-T5 for both CapgMyo DB-b and CapgMyo DB-c).

Table 4.4 Inter-subject gesture recognition accuracies. The average recognition accuracies (%) of 8 hand gestures for CapgMyo DB-b and 12 hand gestures for CapgMyo DB-c for 10 different subjects respectively. The numbers are the majority voted results using 150 ms window (i.e., 150 frames).

Methods	CapgMyo DB-b				
	Number of available trials for adaptation				
	T1	T2	T3	T4	T5
Du et. al. [21],[26]	71.81	86.52	88.66	90.32	91.36
2SRNN [57]	-	-	-	-	89.90
All-ConvNet+TL (Proposed)	<b>75.34</b>	<b>89.42</b>	<b>92.09</b>	<b>93.83</b>	<b>94.94</b>
	CapgMyo DB-c				
Du et. al. [21],[26]	57.40	75.98	82.51	85.98	88.02
2SRNN [57]	-	-	-	-	85.40
All-ConvNet+TL (Proposed)	<b>58.47</b>	<b>78.89</b>	<b>86.02</b>	<b>89.99</b>	<b>91.57</b>

Similar to the inter-session scenario, we also compared our proposed lightweight All-ConvNet+TL in the inter-subject scenario with the state-of-the-art GengNet [21], [26] in a data-starved condition. The proposed lightweight All-ConvNet+TL exhibits improvement over the state-of-the-art on CapgMyo DB-b and CapgMyo DB-c datasets when a limited number of trials are available for adaptation, as observed in Table 4.4, specifically in Columns T1, T2, T3, and T4, respectively. For example, when only 20% of the data (i.e., 1 trial) was available for adaptation, the proposed lightweight All-ConvNet+TL achieved a

3.53% and 1.07% improvement over GengNet [21], [26] on CapgMyo DB-b and CapgMyo DB-c, respectively (Table 4.4, Column- T1).

We summarise the inter-session and inter-subject improvement results in Table 4.5 over the state-of-the-art methods. As indicated there, the performance of the proposed lightweight All-ConvNet+TL is superior in all cases. The improvement achieved by the lightweight All-ConvNet+TL leveraging transfer learning in inter-session and inter-subject scenarios, exceeds those obtained through alternative state-of-the-art domain adaptation approaches.

Table 4.5. Inter-session and Inter-subject improvement (%) results obtained by the proposed lightweight All-ConvNet+TL leveraging transfer learning.

Methods	Inter-session improvement	Inter-subject Improvement	
	DB-b	DB-b	DB-c
Du et. al. [21][26]	6.43	3.58	3.55
2SRNN [57]	<b>11.11</b>	<b>5.04</b>	<b>6.17</b>

Finally, we evaluate the performance of our proposed lightweight All-ConvNet+TL while freezing its maximum number of layers and use them as a feature extractor, and only the top convolutions layers are fine-tuned in the adaptation stage for inter-session and inter-subject gesture recognition. More explicitly, the first six (6) convolutional layers of the lightweight All-ConvNet+TL were frozen and used as a *feature extractor*. Only the top two convolutional layers with a few parameters were fine-tuned in the adaptation stage. Therefore, these experiments can be considered as a full feature extraction setting. The performance of these full feature extraction settings was compared with the more complex computationally expensive 2SRNN [57] method. A deep-stacked RNN classifier was also used as a feature extractor by freezing its weight in the domain adaptation stage. Table 4.6



presents the inter-session and inter-subject average gesture recognition accuracies (%) of 8 and 12 hand gestures for CapgMyo DB-b and DB-c for 10 subjects, respectively. As can be seen from Table 4.6, our proposed lightweight All-ConvNet+TL clearly outperforms the 2SRNN [57] in both *inter-session* and *inter-subject* gesture recognition accuracy. These experimental results indicate that the proposed lightweight All-ConvNet+TL is very effective for discriminative feature extraction for improved gesture recognition in both inter-session and inter-subject scenarios.

Table 4.6 Inter-session and Inter-subject gesture recognition accuracies (%) under full feature extraction setting.

Methods	Inter-session	Inter-subject	
	DB-b	DB-b	DB-c
2SRNN [57]	83.80	89.90	85.40
All-ConvNet+TL (Proposed)	<b>91.93</b>	<b>91.56</b>	<b>85.56</b>

#### 4.6.4 Weight (or Feature) Transfusion Experiments

In this section, we investigate to identify locations where exactly in the network meaningful feature reuse takes place during transfer learning by conducting a weight (or feature) transfusion experiment. We initialize our proposed lightweight All-ConvNet+TL with a contiguous subset of the layers using pre-trained weights (weight transfusion), and the rest of the network randomly, and train on the target inter-session gesture recognition task. More explicitly, we initialize only up to layer  $L$  with pretrained lightweight All-ConvNet+TL weights, and layer  $L + 1$  onwards randomly; then train only layers  $L + 1$  onwards. Since, the weight transfusion process uses pre-trained weights, it can accelerate

the training during fine-tuning of a network on the target task. Therefore, the learning speed was measured in terms of gesture recognition performance on various training epochs. Table 4.7 presents the inter-session gesture recognition accuracy of a subject against various training epochs for different number of transfused weights. We show the learning speed and gesture recognition accuracy when transfusing from Conv1 (L-7, one layer) up to Conv8 (i.e., layer L-7 to layers L-full transfer). From the weight transfusion results, our proposed lightweight All-ConvNet+TL model perform quite stably over the different number of transfused weights. However, we observed that reusing the lowest layers (transfusing weights) leads to the greatest gain in learning speed and gesture recognition accuracy. For example, transfusing weights from layer L-7 (Conv1) up to layer L-5 (Conv3), we achieve  $\approx 98\%$  recognition accuracy after just 8 (eight) training epochs.

#### **4.6.5 Lightweight All-ConvNet Network Trimming**

These weight transfusion results in section 4.6.5 motivate us to explore hybrid approaches to transfer learning, thereby, we introduce network trimming which further optimizes the proposed lightweight All-ConvNet+TL by pruning the weights of the network. We consider reusing pre-trained weights up to Conv3 (i.e., weights of layers L-7 to layers L-5 showed in Table 4.7) and the weights of the top of the lightweight All-ConvNet (i.e., from layers Conv4 (L-4) to Conv7 (L-1)) was pruned by halves to be even more lightweight and initializing these layers randomly. Finally, this new Lightweight All-ConvNet-Slim model was trained or fine-tuned on the target inter-session gesture recognition task. Table 4.8 presents the inter-session gesture recognition accuracy of a subject against various training epochs, which compares the performance of Lightweight All-ConvNet+TL vs Lightweight All-ConvNet-Slim model.

Table 4.7. Learning (or convergence) speed using various training epochs. Table shows inter-session gesture recognition accuracies (%) on test set. The numbers are the majority voted results using 150 ms window (i.e., 150 frames). Per-frame accuracies are shown in parenthesis.

Weight transfusion (up to layers)	Training epochs					
	8	16	32	46	64	100
Full Transfer (L)	70.90 (64.56)	81.74 (67.84)	83.20 (68.35)	83.08 (68.33)	83.21 (68.47)	<b>83.60</b> (68.52)
L-1	87.42 (72.28)	88.21 (73.53)	90.14 (74.43)	90.01 (74.55)	89.85 (74.94)	<b>90.39</b> (75.13)
L-2	90.24 (76.35)	93.60 (78.17)	93.94 (79.62)	94.22 (80.08)	<b>94.50</b> (80.47)	94.18 (81.36)
L-3	95.01 (79.48)	95.96 (81.53)	96.42 (83.23)	96.71 (83.22)	96.99 (83.97)	<b>98.28</b> (84.67)
L-4	96.10 (81.87)	97.71 (82.59)	98.21 (85.10)	97.92 (86.17)	97.96 (86.37)	<b>98.59</b> (87.06)
L-5	97.96 (83.14)	98.40 (84.888)	99.12 (87.00)	99.12 (86.99)	99.28 (87.86)	<b>99.35</b> (88.30)
L-6	98.34 (82.93)	97.76 (85.48)	99.26 (87.24)	98.85 (87.56)	<b>99.27</b> (87.79)	99.25 (88.68)
L-7	98.10 (83.33)	98.74 (84.34)	98.93 (86.08)	<b>99.41</b> (87.22)	99.32 (88.04)	99.32 (88.21)

Table 4.8. Learning (or convergence) speed using various training epochs. Table shows inter-session gesture recognition accuracies (%) on test set. The numbers are the majority voted results using 150 ms window (i.e., 150 frames). Per-frame accuracies are shown in parenthesis.

Model	# learning parameters	Training epochs			
		8	16	24	32
Lightweight All-ConvNet+TL (Proposed)	$\approx 0.46 M$	<b>96.00</b> (71.56)	96.60 (74.79)	97.60 (76.92)	97.69 (77.68)
Lightweight All-ConvNet-Slim (Proposed)	$\approx 0.19 M$	91.92 (68.98)	<b>96.90</b> (73.70)	<b>98.28</b> (75.98)	<b>98.50</b> (77.47)

The experimental results demonstrates that the lightweight All-ConvNet-Slim model can maintain the same or achieve higher performance with much smaller number of parameters. These results with variants of Lightweight All-ConvNet+TL model also highlight many new, rich and flexible ways to use transfer learning.

#### **4.7 Discussion**

We address the problem of distribution shifts by adapting a lightweight model to new target domain tasks using a limited amount of data for sEMG-based inter-session and inter-subject gesture recognition. We propose All-ConvNet+TL leveraging lightweight All-ConvNet and transfer learning, which can be seen as a hybrid of feature extraction and fine-tuning, learning parameters that are discriminative for the new target task. We show the effectiveness of our method by conducting extensive experiments on four (4) publicly available HD-sEMG datasets for three (3) different sEMG-based gesture recognition tasks, including intra-session, inter-session, and inter-subject scenarios. The results indicate that our proposed lightweight All-ConvNet and All-ConvNet+TL models outperform the more complex state-of-the-art models on various tasks and datasets. In intra-session scenarios, the proposed lightweight All-ConvNet (size of only 0.46 M learning parameters), which consists of a network using nothing, but convolutions and subsampling outperformed the most complex state-of-the-art GengNet [21], [26] (size of 5.6M parameters) on CapgMyo DB-b (Session 1 and Session 2) dataset, respectively and performed on par with or very competitively on CapgMyo DB-a and CapgMyo DB-c, respectively. The high intra-session gesture recognition accuracies of 98.02%, 97.52%, 96.80%, and 95.76% were obtained by the proposed lightweight All-ConvNet using a simple majority voting over the recognition

result of 160 instantaneous images (or frames) for DB-a, DB-b (Session 1 and Session 2) and DB-c, respectively.

For gesture recognition in inter-session and inter-subject scenarios, we apply transfer learning to our proposed lightweight All-ConvNet. Our proposed method All-ConvNet+TL leveraging the lightweight All-ConvNet, and transfer learning outperforms the current state-of-the-art methods by a large margin, both when the data from single trials or multiple trials are available for fine-tuning and adaptation.

We achieved state-of-the-art performance for inter-session and inter-subject scenarios. The inter-session gesture recognition accuracy reached 94.1% on CapgMyo DB-b, which is approximately 11.11% and 6.43% higher than the current state-of-the-art [57] and [21], [26], respectively.

In addition, the inter-subject gesture recognition accuracy reached 94.94% and 91.57% on CapgMyo DB-b and DB-c, respectively, which is 5.04% and 6.17% higher than [57] and 3.58% and 3.55% higher than the [21], [26] respectively. Moreover, the proposed lightweight models achieved state-of-art performance under full feature extraction settings in both inter-session and inter-subject scenarios.

These outstanding state-of-the-art inter-session and inter-subject gesture recognition performance achieved by the proposed lightweight All-ConvNet+TL models by leveraging transfer learning validates that the proposed method is highly effective in learning invariant and discriminative representations to overcome the distribution shift caused by inter-session and inter-subject data variability. This potentially indicates that the current state-of-the-art models are overparameterized for the sEMG-based gesture recognition problem.

Furthermore, the current most complex state-of-the-art models [21], [26], [57] are computationally expensive and require a huge memory space to store a massive number of parameters. Therefore, these models are usually unsuitable for deploying low-end, resource-constrained embedded, mobile and wearable devices for real-time MCI applications. Thanks to the proposed parameter-efficient All-ConvNet and All-ConvNet+TL, our model is much smaller and lightweight than these current state-of-the-art methods for sEMG-based gesture recognition.

Finally, the new experimental evidence of our proposed method about various sEMG-based gesture recognition tasks and its role will shed light on potential future directions for the community to move forward for more efficient lightweight model exploration.

#### **4.8 Conclusion**

For real-time Muscle-Computer Interfaces, the sEMG-based gesture recognition must address the inter-session and inter-subject distribution shifts. To address and overcome these distribution shifts, we investigate the effects of transfer learning and feature reuse on our proposed lightweight All-ConvNet. We discovered that the proposed lightweight All-ConvNet+TL, which leverages transfer learning in the inter-session and inter-subject scenarios outperforms the most complex state-of-the-art domain adaptation methods by a large margin, both when the data from single trials or multiple trials are available for adaptation. The state-of-the-art performance proved that the proposed lightweight All-ConvNet+TL model is highly effective in learning invariant and discriminative representations for addressing distribution shifts in sEMG-based inter-session and inter-subject gesture recognition. This raises the question and provides evidence of overparameterization of the most complex current state-of-the-art models for sEMG-based

gesture recognition tasks. We also find that significant feature reuse concentrated in lower layers and explored more flexible and hybrid transfer approaches, which retain transfer benefits and create new possibilities. In future work, we plan to deploy our proposed lightweight All-ConvNet and All-ConvNet+TL model for sEMG-based real-time adaptive and intuitive control of an active prosthesis.

Furthermore, the existing methods (e.g., [21], [26]) achieve only high gesture recognition accuracy based on instantaneous HD-sEMG signals using CNN/DNN methods while reported very low gesture recognition accuracy (e.g., as low as 20%) using classical machine learning methods such as SVM. However, in another study, we argued that SVM can also perform competitively to the more complex state-of-the-art CNN/DNN methods for instantaneous HD-sEMG-based gesture recognition tasks if well-behaved and discriminative features are provided to it. Hence, the next chapter presents a discriminative feature extraction method based on Histogram of Oriented Gradients (HoG) for instantaneous HD-sEMG image recognition, adopting pairwise SVM as the classification scheme, thereby providing a more efficient and an alternative solution to the more complex CNN/DNN methods.

## **Chapitre 5 - HOG and Pairwise SVMs for Neuromuscular Activity Recognition Using Instantaneous HD-sEMG Images**

The concept of neuromuscular activity recognition using instantaneous high-density surface electromyography (HD-sEMG) image opens up new avenues for the development of more fluid and natural muscle-computer interfaces. The state-of-the-art methods for instantaneous HD-sEMG image recognition achieve prominent performance using a computationally intensive deep convolutional networks (ConvNet) classifier, while very low performance is reported using the conventional classifiers. However, the conventional classifiers such as Support Vector Machines (SVM) can surpass ConvNet at producing optimal classification if well-behaved feature vectors are provided. This chapter studies the question of extracting distinctive feature sets, thus propose to use Histograms of Oriented Gradient (HOG) as unique features for robust neuromuscular activity recognition, adopting pairwise SVMs as the classification scheme. The experimental results proved that the HOG represents unique features inside the instantaneous HD-sEMG image and fine-tuning the hyper-parameter of the pairwise SVMs, the recognition accuracy comparable to the more complex state-of-the-art methods can be achieved.



## 5.1 Introduction

The precise characterization and recognition of neuromuscular activities present a great challenge [1]. The current state-of-the-art methods [21], [26] employed a computational model based on deep convolutional neural networks (ConvNet) [35] for sEMG image classification. However, the potential drawback is the classification method based on ConvNet, is computationally very expensive to be practical for real-world applications for gesture or neuromuscular activity recognition. Moreover, the studies conducted in [21], [26] reported of attaining recognition rate as low as 20% using the conventional classifiers such as support vector machines (SVM). However, the conventional classifiers such as SVM can surpass ConvNet at producing optimal classification if well-behaved feature vectors are provided [106]. However, this aspect is totally overlooked in [21], [26]. Therefore, developing computationally efficient distinctive feature extraction and classification algorithms for instantaneous sEMG image based neuromuscular activity recognition is highly demanded.

For instantaneous sEMG image based neuromuscular activity recognition, the challenge remains open because very limited research has been done on it. This chapter studies the histogram of oriented gradients (HOG) for the improved characterization of the instantaneous sEMG image. HOG is one of the state-of-the-art methods for object recognition [107]-[110]. However, this important characterization method is ignored for sEMG signal classification. In this thesis, we propose to use a HOG based feature extraction method for instantaneous sEMG image classification. According to our best knowledge, no one performed similar studies before for sEMG signal classification.

The rest of the chapter is organized as follows. Section 5.2 provides the computational details of the proposed feature extraction method. Section 5.3 describes the testing database and the experimental validation. Section 5.4 offers some conclusive remarks.

## 5.2 The Proposed Neuromuscular Feature Extraction and Classification Algorithm

The proposed neuromuscular feature extraction and classification algorithm has three computational components: (i) preprocessing and sEMG image generation, (ii) feature extraction, and (iii) classification. A schematic diagram of the proposed muscular activity recognition method by instantaneous sEMG images is shown in Fig. 5.1. First, the acquired HD-sEMG signals at each sampling instant were arranged in a 2-D grid according to their electrode positioning. This grid was further transformed into an instantaneous sEMG image by linearly transforming the values of sEMG signals from  $mV$  to color intensity as  $[-2.5mV, 2.5mV]$  to  $[0 255]$ . Thus, an instantaneous grayscale sEMG image was formed with the size of  $16 \times 8$ . The gradient image  $\nabla f(x,y)$  is obtained by convolving an estimation filter over  $x$  and  $y$  axis of the instantaneous sEMG image  $f(x,y)$ . The magnitude  $|\nabla f(x,y)|$  and orientation  $\theta(x,y)$  for each pixel of the sEMG image is computed from the gradient image  $\nabla f(x,y)$ . The sEMG image is divided into a dense grid with a spatial  $\eta \times \eta$  pixels cells. For each cell, a local 1-D histogram of gradient over all pixels in the cell is computed as features. This aggregated cell-level 1-D histogram builds the HOG feature vector for the unique representation of the instantaneous sEMG image. Finally, these HOG feature vectors are fed to a computationally effective learned pairwise SVM classifier for instantaneous gesture recognition.

Section 5.2.1 presents the HOG feature extraction technique for sEMG image representation and Section 5.2.2 presents the classification schemes respectively.

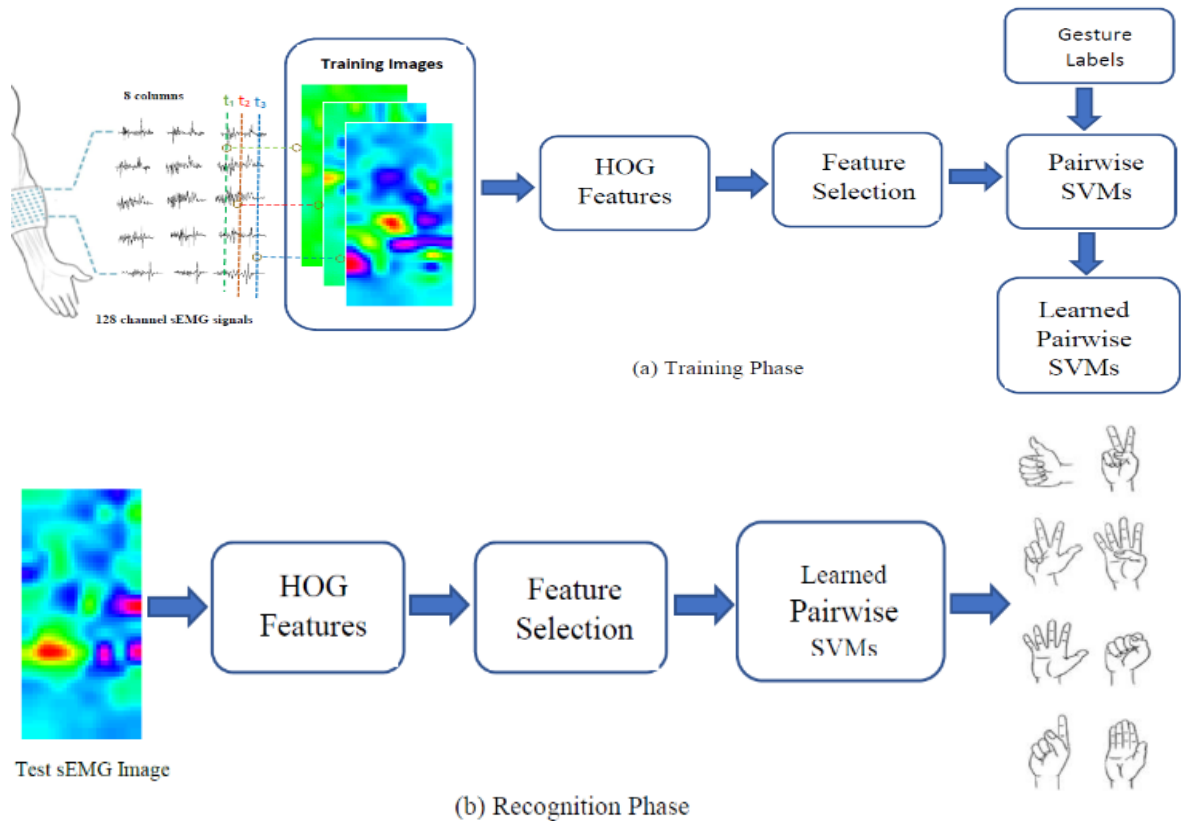


Fig. 5.1 Schematic illustration of the proposed muscular activity recognition by instantaneous sEMG images.

### 5.2.1 Histogram of Oriented Gradients (HOG) Feature Extraction

After generating the instantaneous sEMG image by linearly transforming the values of sEMG signals to color intensity as mentioned above, the crucial task is to extract distinctive features to represent the instantaneous sEMG image for robust classification of the performed hand gesture. However, the main research question is *what makes the different gestures distinctive performed by the same or different subjects?* For example, the hand gestures explained in Chapter 2 in Table 2.1 and Table 2.2 respectively can be differentiated by their shape and orientation features. The color might not be a reliable feature because the portrayed hand gestures have the same color. Therefore, any method

that can precisely describe the shape and orientation information will solve the problem. Nevertheless, the problem in our hand is even more challenging because the instantaneous sEMG image is formed by linearly transforming the values of sEMG signals from mV to color intensity which reflects the intensity distributions of the performed hand gestures. The different hand gestures produce different spatial intensity distributions, thus also make the structure of the instantaneous sEMG image different. These discriminative attributes have been capitalized and used as features in this work.

Both intuitive observation and preliminary experimental results indicate that the gradient of the intensity distributions or edge directions provides the discriminative features for instantaneous sEMG image classification. HOG precisely captures this notion. Therefore, we propose to use HOG as features for instantaneous sEMG image classification. HOG features are calculated by taking orientation histograms of intensity distributions from all locations of a dense grid on a sEMG image region and combined features are used for classification. HOG features are assumed to be designed for imitating the visual information processing of the brain and have robustness against local changes of position. This important property of HOG can be exploited to cope with the electrode shifting problem encountered between two different HD-sEMG recording sessions. HOG is like scale-invariant feature transform [110] in the sense that a local region is described by deriving gradient orientations from the orientation histogram.

Consider the gradient estimation filters  $h_x = [-1, 0, 1]$ , and  $h_y = [-1, 0, 1]^T$ . The gradient information of an instantaneous sEMG image can be obtained by

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T = \begin{bmatrix} f(x, y) * h_x \\ f(x, y) * h_y \end{bmatrix} \quad (5.1)$$

where,  $*$  denotes an operation of a 1-dimensional (1-D) convolution. The  $x$  and  $y$  stand for height and width of the instantaneous sEMG image. The magnitude of a pixel is calculated by

$$|\nabla f(x, y)| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2} \quad (5.2)$$

and the orientation of a pixel is calculated by

$$\theta(x, y) = \tan^{-1}\left(\frac{\partial f}{\partial x} / \frac{\partial f}{\partial y}\right) \quad (5.3)$$

These magnitude  $|\nabla f(x, y)|$  and orientation  $\theta(x, y)$  at each pixel are then used for calculating HOG.

The main intuition behind HOG feature extraction is that, while individual  $|\nabla f(x, y)|$  and  $\theta(x, y)$  are highly variable and subject to significant variations across nearby  $(x, y)$  locations, even for the sEMG images generated by the same hand gesture, the cumulative statistics of the spatial distribution of the gradient orientation and magnitudes over small region of the sEMG images derived from the same gesture provide quite robust descriptors of the instantaneous sEMG image.

To compute orientation histograms, the obtained instantaneous sEMG image gradient is divided into  $8 \times 4 = 32$  non-overlapping rectangular cells, and each cell is of size  $\eta \times \eta$  pixels ( $\eta = 2$ ). Four  $\eta \times \eta$  neighboring cells form a block of size  $\zeta \times \zeta$  ( $\zeta = 2$ ). A schematic diagram of HOG extraction process is illustrated in Fig. 5.2. There are total  $v\zeta \times h\zeta = 21$ , overlapping blocks are formed over an instantaneous sEMG image (where  $v\zeta = 7$  and  $h\zeta = 3$ , denotes the number of vertical and horizontal block respectively). In each  $\eta \times \eta$  cell, the orientation histogram has  $\beta$  bins ( $\beta = 7$ ), which correspond to

orientations  $i \times \pi/\beta$ , where  $i = 0, 1, \dots, \beta$ . Thus, each of the block contains  $\zeta \times \zeta \times \beta = 28$  dimensional HOG feature vectors and each instantaneous sEMG image contains  $v\zeta \times h\zeta \times (\zeta \times \zeta \times \beta) = 588$  dimensional HOG feature vectors.

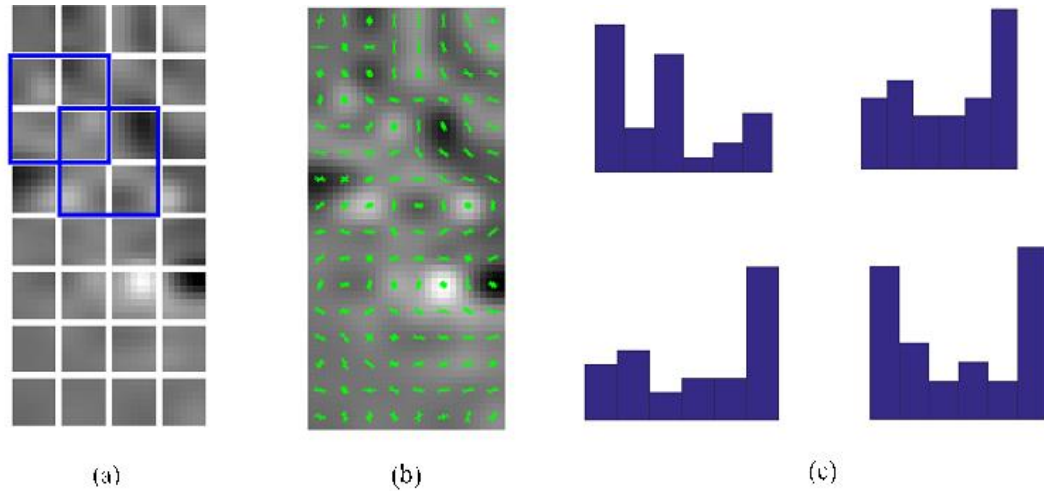


Fig. 5.2 HOG extraction process (a) An instantaneous sEMG image is partitioned by non-overlapping cells and overlapping blocks (each block has  $(2 \times 2)$  four cells). (b) Gradients information are overlaid over an instantaneous sEMG image (c) HOG in each block. The horizontal axis represents angle information and the vertical axis bears weighted histogram.

This 588-dimensional HOG feature vector is used to represent the instantaneous sEMG image. It is noteworthy that  $\eta$ ,  $\zeta$  and  $\beta$  are parameters and selecting values of these parameter tradeoff with the overall instantaneous sEMG image classification performance. Therefore, it is significant to select the optimum values of these parameters for extracting the most discriminant HOG features.

Now, we calculate the 28-D HOG feature vector from a block of  $\zeta \times \zeta$  cells. Consider  $|\nabla f(x, y)|$  and  $\theta(x, y)$  in one block as shown in Fig. 5.2(a) and 5.2(b). In Fig. 5.2(b), the orientation of the arrow represents  $\theta(x, y)$  and the length of the arrow stands for

$|\nabla f(x, y)|$ . In the experiments, the gradient orientation is transformed from  $-\pi \leq \theta \leq \pi$  to  $0 \leq \theta \leq \pi$  and then evenly quantized into  $\beta$  bins.

The HOG feature vector  $h_1 \in \mathbb{R}^\beta$  of the first cell (top left in Fig. 5.2(a)) can be calculated by voting

$$h_1(i) \leftarrow h_1(i) + |\nabla f \theta_i(x, y)|, \quad i = 1, \dots, \beta \quad (5.4)$$

where  $|\nabla f \theta_i(x, y)|$  indicates the magnitude from the gradient and  $\theta_i$  is the quantized orientation. In the same way as  $h_1$ , the three-feature vectors ( $h_2, h_3$  and  $h_4$ ) can be generated from three other cells of a same block. By combining these feature vectors, the HOG feature vectors of a block turn into  $h = [h_1^T, h_2^T, h_3^T, h_4^T]^T \in \mathbb{R}^{\beta \times 4}$ .

It is to be noted that the equation (5.4) is a simplified form. However, in our implementation, trilinear interpolation is used to calculate the HOG features [111]. The trilinear interpolation smoothly distributes the gradient to  $\zeta \times \zeta$  cells of a block to reduce the aliasing effect caused by the pixels near to the cell boundaries. This technique can also be robust against small distortions between sEMG images derived from the same gesture.

Moreover, the gradient strengths vary over an instantaneous sEMG image owing to local variations. Therefore, the overlapped blocks on sEMG image are normalized individually so that each scalar cell-response contributes several components to final HOG feature vector. The normalization is performed by

$$h = h / \sqrt{\|h\|_2^2 + \epsilon^2} \quad (5.5)$$

where,  $\epsilon$  is a small normalization constant used to avoid divided by zero [111]. This normalized HOG representation is used for instantaneous sEMG image classification.

### 5.2.2 Pairwise SVM Classifier

After the HOG feature extraction for representing an instantaneous sEMG image, the most important task is to employ a computationally effective classifier which has the high generalization ability for solving a multi-class classification problem. SVM [112], [113] is essentially a binary classifier, however, multi-class classification problem is solved by training several binary SVM classifiers and an optimal global decision function is obtained by fusing the outputs of each of these binary classifiers. In addition, the decision function of SVM's is fully determined by the number of support vectors (SVs) which is substantially lower than the actual number of samples used in training, makes SVM computationally very efficient. Moreover, SVM trained on HOG features has become a popular method for across many visual perception tasks due to the performance and robust theory [114]. Why do SVM's trained on HOG features perform so well is still an open research issue in the literature. However, it is pointed out in [114] that preserving second-order statistics and locality of interactions are fundamental to achieve good performance. All these motivated us to use and train pairwise SVM's classifiers on HOG features extracted from the instantaneous sEMG image.

## 5.3 Experiments

We tested our feature characterization method on CapgMyo data sets [26] as discussed in Chapter 2, Section 2.5. The CapgMyo database comprises 3 sub-databases (referred to as DB-a, DB-b and DB-c). However, as followed by the [21], DB-a has been used in our preliminary experiments to evaluate the performance of our proposed methods. In DB-a, 8 isotonic and isometric hand gestures were obtained from 18 of the 23 subjects and each gesture was also recorded for 10 times. For each subject, the recorded HD-sEMG data is



filtered, sampled and the instantaneous sEMG image is generated using the method mentioned in section 5.2. More explicitly, 8 different hand gestures are performed by every subject and each hand gestures are recorded for 10 times with a 1000 Hz sampling rate, which in total generates  $(8 \times 10 \times 1000 = 80000)$  instantaneous sEMG images. Then, our HOG-based proposed feature extraction technique elaborated in Section 5.2.1 is applied to each of the instantaneous sEMG images. Thus, an  $80000 \times M$  dimension HOG feature vectors are obtained. The each of the HOG feature vectors dimension  $M$  depend on the different HOG parameters such as  $\eta, \zeta$  and  $\beta$ . However, considering the low resolution instantaneous sEMG image and based on our preliminary experiments, we select  $\eta = 2$ ,  $\zeta = 2$  and  $\beta = 7$  respectively. Hence, we obtained  $v\zeta \times h\zeta \times (\zeta \times \zeta \times \beta) = 588$  dimension HOG feature vectors of an instantaneous sEMG image.

Now, for every subject in DB-a, a pairwise SVMs classifier is trained to predict the desired hand gestures for each incoming sEMG images. The pairwise SVMs framework is based on LIBSVM, a library for support vector machines [115]. To conduct the above-mentioned gesture classification task, the obtained  $80000 \times M$  dimension HOG feature vectors are divided into three subsets such as training, validation and testing set. In this initial investigation, 50% of the Histogram of Oriented Gradients (HOG) feature vectors from the complete feature set, which corresponds to half of the trials (i.e., 5 trials for each specific movement), are chosen and employed as the training set. Similarly, the remaining 50% of the HOG feature vectors, which correspond to the other half (5 trials), are further split into a validation set and a testing set. The validation set is used for model/kernel and parameter selection for pairwise SVMs. Due to computationally effective and reducing searching space for parameter selection, the RBF kernel  $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$ ,  $\gamma > 0$  is used to

train the obtained HOG feature set. There are two parameters for an RBF kernel which is a cost parameter ( $C$ ) and kernel parameter  $\gamma$ . It is not known in advance which  $C$  and  $\gamma$  are the best for a given problem. Therefore, the parameter selection is performed. We used a grid search along with this  $v$ -fold ( $v = 3$ ) cross-validation scheme to find the optimum  $(C, \gamma)$  on the validation set. It is recommended in [116] to use the exponentially growing sequences of  $C$  and  $\gamma$  to identify the good parameters. Hence, we use  $C = [2^5, 2^4, 2^3, \dots, 2^{-1}]$  and  $\gamma = [2^{-6}, 2^{-5}, 2^{-4}, \dots, 2^2]$ . Therefore, we examined with  $7 \times 9 = 63$  combinations of  $(C, \gamma)$  pairs. Then, the whole training feature set is trained using the pair of  $(C, \gamma)$  that achieves the best cross-validation accuracy. Finally, this trained classifier is used to predict the test feature set.

Confusion matrix generated from the predicted classification results were used as a performance indicator. The correctly classified (%) gesture classes are listed along the diagonal line of the Confusion matrix as presented in Fig. 5.3. The average classification accuracy of the proposed methods is 86.63% which is comparable to the state-of-the-art methods. Using instantaneous values of HD-sEMG and SVM classifier, the average classification accuracy as low as 20% was reported in [21]. However, the average classification accuracy increased to 86.63% using proposed HOG and optimized parameter of pairwise SVMs. In addition, the recall or true positive rate (TPR) and the precision or the positive predictive value (PPV) [117] of each gesture classes are also computed and mentioned in Table 5.1. The 86.62% average precision and recall of each class also indicate the potentiality of the HOG and pairwise SVMs for neuromuscular activity recognition. Finally, the experimental results demonstrate that: (i) HOG are effective features for unique representations of instantaneous HD-sEMG images (ii) Provided discriminant features and

fine-tuning the hyper-parameter of the conventional classifiers such as pairwise SVMs, the state-of-the-art recognition rate can be achieved for muscular activity recognition based on instantaneous HD-sEMG images.

True Class	CL01	<b>87.35</b>	0.39	1.05	0.04	0.08	7.85	3.24	0.00
	CL02	0.40	<b>86.03</b>	4.72	4.44	2.06	0.61	0.12	1.61
	CL03	1.38	3.88	<b>89.24</b>	1.74	1.29	0.44	0.89	1.13
	CL04	0.04	3.51	1.24	<b>82.84</b>	4.51	0.60	0.40	6.86
	CL05	0.04	1.76	1.12	4.31	<b>89.99</b>	0.28	0.08	2.43
	CL06	8.77	0.57	0.89	0.53	0.81	<b>83.76</b>	4.14	0.53
	CL07	2.27	0.20	1.34	0.20	0.24	4.05	<b>89.02</b>	2.67
	CL08	0.04	1.93	1.26	6.52	2.63	0.51	2.32	<b>84.79</b>
		CL01	CL02	CL03	CL04	CL05	CL06	CL07	CL08
	<b>Predicted Class</b>								

Fig. 5.3 Confusion Matrix of the Proposed Neuromuscular Activity Recognition Method.

Table 5. 1 Precision and Recall of every gesture classes.

Class	CL01	CL02	CL03	CL04	CL05	CL06	CL07	CL08
<b>Precision</b>	87.52	87.44	88.38	82.32	88.57	85.07	88.66	85.03
<b>Recall</b>	87.35	86.03	89.24	82.84	89.99	83.76	89.02	84.79

## 5.4 Conclusions

In this Chapter, we propose to use Histogram of Oriented Gradients (HOGs) as distinctive features and pairwise SVMs for robust neuromuscular activity recognition using instantaneous HD-sEMG images. 80000 instantaneous HD-sEMG image frames for 8 different gestures of each subject from CapgMyo database were examined. The experimental results demonstrate that HOG are effective features for unique representations

of instantaneous HD-sEMG images. Also, provided discriminant features and fine-tuning the hyper-parameter of the conventional classifiers such as pairwise SVMs, the state-of-the-art recognition rate can be achieved for neuromuscular activity recognition based on instantaneous HD-sEMG images.

# Chapitre 6 - Conclusion

## 6.1 Summary

The need for sufficient amount of labeled data, high-end computational resources and the presence of distribution shift in inter-session and inter-subject scenarios are the major factors that impede deploying deep learning for real-time sEMG-based gesture recognition tasks. There is a significant demand for cost-effective, compact and lightweight models that not only effectively address these issues but also achieve competitive performance in high-end resource constrained scenarios when compared to the more complex current state-of-the-art models. The current state-of-the-art models have upwards of  $> 6M$  parameters to learn.

In this thesis, we first present low-latency S-ConvNet, a simple yet efficient framework for learning instantaneous HD-sEMG images from scratch through random initialization for gesture or neuromuscular activity recognition. The experimental results proved that the proposed S-ConvNet is very effective for learning discriminative muscle activation features for instantaneous HD-sEMG image recognition especially in the data and high-end resource constrained scenarios. Without using any pre-trained models, our proposed S-ConvNet demonstrate state-of-the-art performance on three (3) out of four (4) publicly available benchmark HD-sEMG datasets, while using  $\approx 12\times$ smaller training dataset and reducing

learning parameters to only  $\approx 2\text{M}$  for sEMG-based gesture recognition in intra-session scenarios outperforming the more complex state-of-the-art.

In addition, electrode shifts, rotations and malfunctions are the serious challenges in sEMG-based gesture recognition. To address these issues and devise more efficient and lightweight network models, the All-ConvNet is introduced. Comprising solely of convolutional layers, this architecture offers a simple yet effective framework for learning instantaneous HD-sEMG images from scratch via random initialization. The inherent design of All-ConvNet's convolutional layers ensures scale and shift invariance, ensuring the model's robustness to variations in signal distributions or amplitude and temporal alignment. This is crucial for accurate recognition across diverse movement scenarios. Similar to the proposed S-ConvNet, without using any pre-trained models, the lightweight All-ConvNet achieve state-of-the-art performance on three (3) publicly available benchmarks HD-sEMG datasets and perform very competitively to the most complex state-of-the-art methods on another compared benchmark HD-sEMG dataset for intra-session gesture recognition based on instantaneous values of HD-sEMG signals. Notably, the proposed All-ConvNet accomplished this remarkable state-of-the-art intra-session gesture recognition performance while operating with an approximately  $\approx 12\times$  smaller dataset and reducing the number of training parameters to only  $\approx 460\text{k}$ .

These exceptional state-of-the-art experimental results for sEMG-based gesture recognition in intra-session scenarios potentially indicate that the proposed low-latency S-ConvNet and the lightweight All-ConvNet are highly effective in learning discriminative feature representations from instantaneous sEMG images. Hence, the proposed S-ConvNet and All-ConvNet models hold significant potential for deploying real-time MCI applications

based on sEMG signals, especially under limited data and high-end resource constrained scenarios.

Moreover, for real-time Muscle-Computer Interfaces, the sEMG-based gesture recognition must address the *inter-session* and *inter-subject* distribution shifts. To further address this distribution shift problem, a domain adaptation method with shallow convolutional neural network (S-ConvNet) is proposed. The proposed domain adaptation (DA) methods with S-ConvNet effectively transfer learned representations from the source domain sEMG dataset or task to target domain's sEMG-based inter-session and/or inter-subject gesture recognition tasks, thereby providing a solution for achieving strategic knowledge transfer and optimal model adaptation. Experiments conducted for gesture recognition in inter-session and inter-subject scenarios on four (4) publicly available benchmark HD-sEMG datasets, the proposed DA methods with S-ConvNet outperformed the current most complex state-of-the-art DA methods.

We further address the problem of distribution shifts by adapting the proposed lightweight All-ConvNet model to new target domain tasks using a limited amount of data for sEMG-based inter-session and inter-subject gesture recognition. We propose All-ConvNet+TL leveraging lightweight All-ConvNet and transfer learning, which can be seen as a hybrid of feature extraction and fine-tuning, learning parameters that are discriminative for the new target task. However, feature extraction and fine-tuning both have their own limitations. To address these limitations, we introduce and conducted a weight (or feature) transfusion experiment in order to find out where does exactly the feature reuse takes place in the network of the proposed TL framework. We find out that meaningful feature reuse is restricted to the lowest few layers of the network. Building upon the findings of these

weight (or feature) transduction experiments, we introduce a network trimming method to further optimize the proposed lightweight All-ConvNet+TL model. This involves selectively pruning the network's weights, resulting in the development of a more efficient Lightweight All-ConvNet-Slim model. The proposed lightweight All-ConvNet-Slim model can maintain the same level of gesture recognition performance or potentially achieve an even higher level, all while requiring the learning of a mere  $\approx 190k$  parameters. Experiments on four datasets demonstrate that the proposed All-ConvNet+TL methods outperform the most complex existing approaches and achieve state-of-the-art results for sEMG-based gesture recognition in inter-session and inter-subject scenarios.

The proposed DA method with S-ConvNet and the lightweight All-ConvNet+TL both sets a new state-of-the-art performance on all four (4) benchmark HD-sEMG dataset outperforming the current state-of-the-art DA methods. The performance gap with the best existing DA methods even increases more when the tiny amount of data (e.g., single trials) were available for adaptation. This state-of-the-art performance proved that the proposed DA with S-ConvNet and the lightweight All-ConvNet+TL model is highly effective in learning domain-invariant and discriminative representations for addressing distribution shifts in sEMG-based gesture recognition in inter-session and inter-subject scenarios. These outstanding experimental results provide evidence that the current state-of-the-art models may be overparameterized for sEMG-based inter-session and inter-subject gesture recognition tasks.

Moreover, deep learning methods require a substantial amount of labeled sEMG data and high computational resources for achieving acceptable gesture recognition accuracy for MCI based on sEMG signals. In addition to these issues, the existing approaches for



sEMG-based gesture recognition using instantaneous sEMG images reported very low recognition accuracy with the classical machine learning methods such as SVM. To address these issues and propose an alternative competitive solution, in another study, we argued that the SVM can perform competitively to the more complex state-of-the-art deep learning methods if well-behaved distinctive features are provided to it. Therefore, we propose to use Histogram of Oriented Gradients (HOGs) as distinctive features for robust gestures or neuromuscular activity recognition using pairwise SVMs as the classification scheme. The experimental results demonstrate that HOG are effective features for unique representations of instantaneous HD-sEMG images and fine-tuning the hyper-parameter of the classical machine learning methods such as SVM, the very competitive gesture recognition performance can be achieved to the more complex state-of-the-art deep learning methods. The proposed method based on HOG and pairwise SVMs also can be effectively deployed to the data constrained and resource bounded scenarios.

## **6.2 Future work and directions**

This thesis has covered various issues in sEMG-based intra-session, inter-session, and inter-subject gesture recognition under limited data availability and resource-constrained scenarios. The need for computationally and memory-efficient methods in sEMG-based MCIs arises from the practical requirements of real-time processing, accommodating limited hardware resources, optimizing energy consumption, providing a seamless user experience, and facilitating deployment in various settings. The proposed methods effectively address these issues while preserving state-of-the-art performance. On the basis of this thesis there are several potential research directions:

- sEMG-based MCIs are often used in portable, mobile, or wearable devices, which may possess limited computational power and memory. Furthermore, these mobile and wearable devices used in MCIs are typically powered by batteries or other energy sources. Therefore, in order to reduce latency, storage requirements, and energy consumption and to run inference more efficiently on these mobile wearable devices, the proposed S-ConvNet and All-ConvNet models, presented in chapters 3 and 4 respectively, can be further optimized by compressing the network using model compression methods such as network pruning and quantization [118].
- In chapter 4 of this thesis, the All-ConvNet-Slim model, with a size of only 190  $k$  learning parameters, is introduced. An experiment conducted in an inter-session scenario has demonstrated its potential for sEMG-based gesture recognition. Further experimental validation of the All-ConvNet-Slim model using diverse sparse-channel and high-density sEMG (HD-sEMG) datasets, across various tasks including intra-session and inter-subject analyses, could represent a promising avenue for future research. In addition, exploring multi-stream and multi-view representation learning with the proposed All-ConvNet-Slim model presents itself as a viable research direction.
- The proposed knowledge-sharing-based domain adaptation (DA) methods employing S-ConvNet and All-ConvNet+TL have effectively demonstrated their capabilities in providing discriminative and domain-invariant feature representations, resulting in state-of-the-art performance across four publicly available HD-sEMG datasets. This achievement raises an intriguing question: Can state-of-the-art performance in sEMG-based inter-session/inter-subject gesture

recognition be achieved by employing the proposed S-ConvNet and All-ConvNet models to align the statistical distribution shift between the source domain and the target domain sEMG dataset or tasks, using statistical mechanisms such as correlation alignment (CORAL) [120]? This could represent a promising research direction for exploration.

- Another promising avenue for research could involve the development of an adversarial discriminative domain adaptation technique [121], [122] based on the proposed S-ConvNet and All-ConvNet for domain invariant feature representation for enhanced inter-session and inter-subject gesture recognition.
- Current approaches in computer vision applications, especially in image classification and object detection task reported significant gain in performance when the ConvNet is augmented by a self-attention mechanism without any parameter overhead [119]. This is achieved by concatenating the convolutional feature maps with a set of feature maps produced via self-attention. Hence, investigating the feasibility of augmenting the proposed S-ConvNet and All-ConvNet models with a self-attention mechanism holds the potential for valuable exploration for sEMG-based gesture recognition.
- The scalability of the proposed HoG and pairwise SVM classifier to sEMG-based gesture recognition in inter-session and inter-subject scenarios could represent an alternative and viable solution to current state-of-the-art deep learning approaches.
- To integrate our developed state-of-the-art models for sEMG-based gesture recognition to the real-time MCI applications such as controlling a GEN3 robotic arm from Kinova [125].

## References

- [1] D. Farina *et al.*, "The extraction of neural information from the surface EMG for the control of upper-limb prostheses: merging avenues and challenges," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 4, pp. 797–809, Jul. 2014.
- [2] G. Jang, J. Kim, J. S. Lee, and Y. Choi, "EMG-based continuous control scheme with simple classifier for electric-powered wheelchair," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 6, pp.3695–3705, 2016.
- [3] R. Jimenez-Fabian, and O. Verlinden, "Review of control algorithms for robotic ankle systems in lower-limb orthoses, prostheses, and exoskeletons," *Medical Engineering & Physics*, vol. 34, no. 4, pp. 397–408, May 2012.
- [4] O. Marin-Pardo *et al.* "A virtual reality muscle-computer interface for neurorehabilitation in chronic stroke: a pilot study." *Sensors (Basel, Switzerland)* vol. 20,13 3754. 4 Jul. 2020.
- [5] Y. Hu, J. N. Mak, & K. Luk, "Application of surface EMG topography in low back pain rehabilitation assessment," *International IEEE/EMBS Conference on Neural Engineering*, pp. 557–560, May 2007.
- [6] D.-H. Kim, *et al.*, "Epidermal electronics," *Science*, vol. 333, pp.838–843, 2011.
- [7] N. Nadia, S. Orts-Escolano, and M. Cazorla., "An sEMG-controlled 3D game for rehabilitation therapies: real-time time hand gesture recognition using deep learning techniques," *Sensors* 20, no. 22: 6451, 2020.
- [8] T. R. Farrell and R. F. f. Weir, "A comparison of the effects of electrode implantation and targeting on pattern classification accuracy for prosthesis control," in *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 9, pp. 2198-2211, Sept. 2008.
- [9] M.A. Oskoei, and H. Hu. "Support vector machine-based classification scheme for myoelectric control applied to upper limb." *IEEE Transactions on Biomedical Engineering*, 55, pp.1956–1965, 2008.
- [10] Z. Lu, X. Chen, Q. Li, X. Zhang, and P. Zhou. "A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices." *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 293–299, 2014.
- [11] K. Li, J. Zhang, L. Wang, M. Zhang, J. Li, and S. Bao, "A review of the key technologies for sEMG-based human-robot interaction systems," *Biomed. Signal Process. Control* 62 (2020), 102074, <https://doi.org/10.1016/j.bspc.2020.102074>.
- [12] E. Costanza., S. A. Inverso, R. Allen, and P Maes, "Intimate interfaces in action: assessing the usability and subtlety of EMG-based motionless gestures," *Conference on Human Factors in Computing Systems*, ACM, pp. 819–828, 2007.
- [13] T. S. Saponas, D. S. Tan, D. Morris, and R. Balakrishnan, "Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces," *Conference on Human Factors in Computing Systems*, ACM, pp. 515–524, 2008.

- [14] T. S. Saponas, D. S. Tan, D. Morris, D. J. Turner and J. A. Landay, "Making muscle-computer interfaces more practical," *Conference on Human Factors in Computing Systems*, pp. 851–854, ACM, 2010.
- [15] M. Atzori et al., "Electromyography data for non-invasive naturally controlled robotic hand prostheses," *Scientific Data 1*, 2014.
- [16] N. Patricia, T. Tommasi. and B. Caputo, "Multi-source adaptive learning for fast control of prosthetics hand," *International Conference on Pattern Recognition*, pp. 2769–2774, 2014.
- [17] C. Amma, T. Krings, J. Ber, J. and T. Schultz, "Advancing muscle computer interfaces with high-density electromyography," *Conference on Human Factors in Computing Systems*, pp. 929–938, ACM, 2015.
- [18] A. Stango, F. Negro and D. Farina, "Spatial correlation of high-density EMG signals provides features robust to electrode number and shift in pattern recognition for myocontrol," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol 23, no. 2, pp. 189–198, 2015.
- [19] M. Atzori et al., "Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands," *Frontiers Neurorobot.*, vol. 10, pp. 9–18, 2016.
- [20] X. Zhai et al., "Self-recalibrating surface EMG pattern recognition for neuroprosthesis control based on convolutional neural network," *Frontiers Neurosci.*, vol. 11, pp. 379–389, 2017.
- [21] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu and J. Li, "Gesture recognition by instantaneous surface EMG images," *Scientific Reports*, Vol 15, no. 6, 36571, Nov 2016.
- [22] M. R. Islam, D. Massicotte, F. Nougrou and W. Zhu, "HOG and pairwise SVMs for neuromuscular activity recognition using instantaneous HD-sEMG images," *IEEE International New Circuits and Systems Conference (NEWCAS)*, Montreal, QC, 2018, pp. 335-339.
- [23] W.T. Wei, Y.K. Wong, Y. Du, Y. Hua, M. Kankanhalli, and W.D. Geng, "A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface," *Pattern Recognit. Lett.* (2017).
- [24] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, and W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PLoS ONE* 13(10): e0206049. <https://doi.org/10.1371/journal.pone.0206049>, 2018.
- [25] M. R. Islam, D. Massicotte, F. Nougrou, P. Massicotte and W. -P. Zhu, "S-Convnet: a shallow convolutional neural network architecture for neuromuscular activity recognition using instantaneous high-density surface EMG images," *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020, pp. 744-749.
- [26] Y. Du., W. Jin, W. Wei, Y. Hu and W Geng, "Surface EMG based inter-session gesture recognition enhanced by deep domain adaptation," *Sensors*, vol. 17, no. 3, 458, 2017.
- [27] M. R. Islam, D. Massicotte and W. Zhu, "All-ConvNet: A lightweight all CNN for neuromuscular activity recognition using instantaneous high-density surface EMG images", *IEEE Int. Instrum. Meas. Technol. Conf.*, pp. 1-6, 2020.
- [28] F. Nougrou, A. Campeau-Lecours, R. Islam, D. Massicotte and B. Gosselin, "Muscle activity distribution features extracted from HDsEMG to perform forearm pattern recognition," *2018 IEEE Life Sciences Conference (LSC)*, Montreal, QC, pp. 275-278, Oct. 2018.

- [29] Tam, M. Boukadoum, A. Campeau-Lecours and B. Gosselin, "A fully embedded adaptive real-time hand gesture classifier leveraging HD-sEMG and deep learning", *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 2, pp. 232-243, Apr. 2020.
- [30] F. Nougrou, A. Campeau-Lecours, D. Massicotte, M. Boukadoum, C. Gosselin, and B. Gosselin. "Pattern recognition based on HD-sEMG spatial features extraction for an efficient proportional control of a robotic arm." *Biomedical Signal Processing and Control* 53 (2019): 101550.
- [31] U. Côté-Allard et al., "Deep learning for electromyographic hand gesture signal classification using transfer learning," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 760-771, April 2019.
- [32] Y. Zou and L. Cheng, "A transfer learning model for gesture recognition based on the deep features extracted by CNN," in *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 5, pp. 447-458, Oct. 2021, doi: 10.1109/TAI.2021.3098253.
- [33] F. D. Farfan, J. C. Politti, and C. J. Felice, "Evaluation of EMG processing techniques using information theory," *Biomed. Eng. Online*, vol. 9, no. 1, pp. 1–18, 2010.
- [34] A. Krasoulis, S. Vijayakumar, and K. Nazarpour, "Multi-grip classification-based prosthesis control with two EMG-IMU sensors," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 508–518, Feb. 2020.
- [35] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: closing the gap to human-level performance in face verification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701-1708, 23-28 June 2014.
- [36] J. Chen, Jiangcheng, B. Sheng, G. Zhang, and G. Cao., "High-density surface EMG-based gesture recognition using a 3D convolutional neural network," *Sensors* 2020, vol 20, no. 4: 1201. <https://doi.org/10.3390/s20041201>
- [37] L. Yanghao., W. Naiyan, S. Jianping, L. Jiaying and H. Xiaodi, "Revisiting batch normalization for practical domain adaptation," *arXiv:1603.04779*, 2016.
- [38] S. Ji, W. Xu, M. Yang and K. Yu, "3D convolutional neural networks for human action recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan. 2013.
- [39] K. He, R. Girshick, and P. Dollár, "Rethinking imagenet pre-training," *IEEE International Conference on Computer Vision (ICCV)*, Seoul, 2019, pp. 4917-4926.
- [40] Z. Li and D. Hoiem, "Learning without forgetting," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935-2947, 1 Dec. 2018.
- [41] J. Ba and R. Caruana., "Do deep nets really need to be deep?," in *Advances in neural information processing systems (NIPS)*, pages 2654–2662, 2014.
- [42] G. Hinton, O. Vinyals, and J. Dean., "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [43] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.
- [44] W. Park, D. Kim, Y. Lu and M. Cho, "Relational knowledge distillation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3962-3971.
- [45] L. Pang, Y. Lan, J. Xu, J. Guo, and X. Cheng., "Locally smoothed neural networks," *In Proceedings of Machine Learning Research*, 77:177–191, *ACML* 2017.

- [46] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller., “Striving for simplicity: the all convolutional net,” *In ICLR*, 2015. *CoRR*, abs/1412.6806
- [47] M. Lin, Q. Chen, and S. Yan, “Network in network,” *In ICLR: Conference Track*, 10 pages, 2014.
- [48] D.-A. Clevert, T. Unterthiner, and S. Hochreiter., “Fast and accurate deep network learning by exponential linear units (elus),” *arXiv preprint arXiv:1511.07289*, 2015.
- [49] K. Janocha, and W. M. Czarnecki. "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.
- [50] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.
- [51] S. J. Pan and Q. Yang, "A survey on transfer learning," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.
- [52] X. Glorot and Y. Bengio., “Understanding the difficulty of training deep feed forward neural networks,” *In AISTATS*, 2010.
- [53] K. He, X. Zhang, S. Ren, and J. Sun., “Delving deep into rectifiers: surpassing human-level performance on imagenet classification,” *In ICCV*, 2015.
- [54] R. Caruana, S. Lawrence and C. Giles, “Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping,” *NIPS*, 2000.
- [55] S. Ioffe and C. Szegedy., “Batch normalization: accelerating deep network training by reducing internal covariate shift,” *In ICML*, 2015.
- [56] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 1, pp.1929–1958, 2014.
- [57] I. Ketykó, F. Kovács and K. Z. Varga, “Domain adaptation for sEMG-based gesture recognition with recurrent neural networks,” *arXiv:1901.069582019*.
- [58] K-O Cho, H-J Jang, “Comparison of different input modalities and network structures for deep learning-based seizure detection,” *Sci Rep* 10, 122 (2020).
- [59] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, “Language modeling with gated convolutional networks,” *In Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 933–941, 2017.
- [60] E. Scheme and K. Englehart, “Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use,” *J. Rehabil. Res. Develop.*, vol. 48, no. 6, pp. 643–59, 2011, doi: 10.1682/jrrd.2010.09.0177, PMID: 21938652.
- [61] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli and W. Geng, “Surface-electromyography-based gesture recognition by multi-view deep learning,” in *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2964-2973, Oct. 2019.
- [62] M. Raghu, C. Zhang, J. Kleinberg and S. Bengio, “Transfusion: understanding transfer learning for medical imaging”, *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS)*, article no.: 301, Pages 3347–3357, December 2019.

- [63] M. R. Islam, D. Massicotte, F. Nougrou, P. Massicotte and W-P Zhu, "S-ConvNet: A shallow convolutional neural network architecture for neuromuscular activity recognition using instantaneous high-density surface EMG images," *arXiv preprint arXiv:1906.03381*, 2019.
- [64] R. N. Khushaba and K. Nazarpour, "Decoding HD-EMG signals for myoelectric control - how small can the analysis window size be?," in *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8569-8574, Oct. 2021, doi: 10.1109/LRA.2021.3111850.
- [65] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International Conference in Machine Learning (ICML)*, 2014.
- [66] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [67] M. R. Islam, D. Massicotte, P. Massicotte and W-P Zhu, "Surface EMG-based inter-session/inter-subject gesture recognition by leveraging lightweight All-ConvNet and transfer learning," *arXiv preprint, arXiv:2305.08014*, 2023.
- [68] R. M. Rangayyan, "Biomedical Signal Analysis," vol. 33. Hoboken, NJ, USA: Wiley, 2015.
- [69] P. Konrad, *The abc of emg*. [http://www.noraxon.com/sdm\\_downloads/abc-of-emg](http://www.noraxon.com/sdm_downloads/abc-of-emg) (2005).
- [70] Q.-C. Ding, A.-B. Xiong, X.-G. Zhao, and J.-D. Han, "A review on researches and applications of sEMG-based motion intent recognition methods," *Acta Automatica Sinica*, vol. 42, no. 1, pp. 13–25, Jan. 2016.
- [71] C. Fleischer and G. Hommel, "A human--exoskeleton interface utilizing electromyography," in *IEEE Transactions on Robotics*, vol. 24, no. 4, pp. 872-882, Aug. 2008
- [72] M. Hassan, H. Kadone, T. Ueno, Y. Hada, Y. Sankai and K. Suzuki, "Feasibility of synergy-based exoskeleton robot control in hemiplegia," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 6, pp. 1233-1242, June 2018.
- [73] A. Phinyomark, P. Phukpattaranont, & C. Limsakul., "Feature reduction and selection for EMG signal classification." *Expert Systems with Applications*, vol. 39, no. 8, pp. 7420–7431, 2012.
- [74] A. Phinyomark, F. Quaine, S. Charbonnier, C. Serviere, F. Tarpin-Bernard, & Y. Laurillau., "EMG feature evaluation for improving myoelectric pattern recognition robustness." *Expert Systems with Applications*, vol. 40, no. 12, pp. 4832–4840, Sep. 2013.
- [75] J. -U. Chu, I. Moon, Y. -J. Lee, S. -K. Kim and M. -S. Mun, "A supervised feature-projection-based real-time EMG pattern recognition for multifunction myoelectric hand control," in *IEEE/ASME Transactions on Mechatronics*, vol. 12, no. 3, pp. 282-290, June 2007.
- [76] Al-Timemy AH, Bugmann G, Escudero J, Outram N., "Classification of finger movements for the dexterous hand prosthesis control with surface electromyography," *IEEE J Biomed Health Inform.* 2013 May;17(3):608-18.
- [77] Y. Hu, J. N. Mak, & K. Luk. "Application of surface EMG topography in low back pain rehabilitation assessment." In *International IEEE/EMBS Conference on Neural Engineering*, pp. 557–560, May 2007.
- [78] B.-U. Kleine, N.-P. Schumann, D. F. Stegeman, & H.-C. Scholle, "Surface EMG mapping of the human trapezius muscle: the topography of monopolar and bipolar surface EMG amplitude and spectrum parameters at varied forces and in fatigue." *Clinical Neurophysiology*, vol. 111, no. 2, pp.686–693, 2000.



- [79] G. Drost, D.F. Stegeman, B.G.M van Engelen and M.J. Zwarts, "Clinical applications of high-density surface EMG: A systematic review," *Journal of Electromyography and Kinesiology*, vol. 16, no. 6, 586–602, Dec 2006.
- [80] M. Rojas-Martínez, M. Mañanas, J. Alonso, and R. Merletti, "Identification of isometric contractions based on high density EMG maps," *Journal of Electromyography and Kinesiology*, vol. 23, no. 1, pp. 33–42, 2013.
- [81] M. Ison, I. Vujaklija, B. Whitsell, D. Farina, and P. Artemiadis, "High-density electromyography and motor skill learning for robust long-term control of a 7-dof robot arm." *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 4, 424–433, 2015.
- [82] L. Z. Bi, A. G. Feleke, and C. T. Guan, "A review on EMG-based motor intention prediction of continuous human upper limb motion for human-robot collaboration," *Biomed. Signal Process. Control*, vol.51, pp. 113–127, May 2019.
- [83] M. Zecca, S. Micera, MC Carrozza, P. Dario, "Control of multifunctional prosthetic hands by processing the electromyographic signal," *Crit Rev Biomed Eng.* 2002;30(4-6):459-85.
- [84] R. N. Khushaba, A. H. Al-Timemy, A. Al-Ani and A. Al-Jumaily, "A framework of temporal-spatial descriptors-based feature extraction for improved myoelectric pattern recognition," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, no. 10, pp. 1821-1831, Oct. 2017.
- [85] Hudgins B, Parker P, Scott RN., "A new strategy for multifunction myoelectric control,". *IEEE Transactions on Biomedical Engineering.* 1993; 40(1):82–94.
- [86] Du YC, Lin CH, Shyu LY, Chen T., "Portable hand motion classifier for multi-channel surface electromyography recognition using grey relational analysis,". *Expert Syst Appl.* 2010.
- [87] Q. Wu et al., "Classification of EMG Signals by BFA-Optimized GSVCM for Diagnosis of Fatigue Status," in *IEEE Transactions on Automation Science and Engineering*, vol. 14, no. 2, pp. 915-930, April 2017.
- [88] F. Duan et al., "sEMG-based identification of hand motion commands using wavelet neural network combined with discrete wavelet transform," *IEEE Trans. Ind. Electron.*, vol. 63, no. 3, pp. 1923–1934, 2016.
- [89] K. Kiatpanichagij and N. Afzulpurkar, "Use of supervised discretization with PCA in wavelet packet transformation-based surface electromyogram classification," *Biomed Signal Processing and Control*, vol. 4, no. 2, pp. 127–138, 2009.
- [90] J. Kilby and H. G. Hosseini, "Extracting effective features of SEMG using continuous wavelet transform," in *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 1704–1707.
- [91] Samadani AA, Kulic D., "Hand gesture recognition based on surface electromyography," In: *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*; 2014. p. 4196–4199.
- [92] Kim J, Mastnik S, Andre' E., "EMG-based hand gesture recognition for realtime biosignal interfacing," In *International Conference on Intelligent User Interfaces*; 2008. p. 30–39.
- [93] Rojas-Martínez, M., Mañanas, M. A. & Alonso, J. F., "High-density surface EMG maps from upper-arm and forearm muscles," *Journal of Neuroengineering and Rehabilitation* 9, 1 (2012).

- [94] D. Wu, J. Yang and M. Sawan, "Transfer Learning on Electromyography (EMG) Tasks: Approaches and Beyond," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, doi: 10.1109/TNSRE.2023.3295453.
- [95] L. Hargrove, K. Englehart and B. Hudgins, "A training strategy to reduce classification degradation due to electrode displacements in pattern recognition based myoelectric control", *Biomed. Signal Process. Control*, vol. 3, no. 2, pp. 175-180, 2008.
- [96] D. Farina and A. Holobar, "Characterization of Human Motor Units From Surface EMG Decomposition," in *Proceedings of the IEEE*, vol. 104, no. 2, pp. 353-373, Feb. 2016.
- [97] Casale, R. & Rainoldi, A., "Fatigue and fibromyalgia syndrome: clinical and neurophysiologic pattern," *Best Practice & Research Clinical Rheumatology* 25, 241–247, 2011.
- [98] [18] S. Gupta, J. Hoffman, and J. Malik., "Cross modal distillation for supervision transfer," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas 2016, pp. 2827-2836.
- [99] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue., "DSOD: Learning deeply supervised object detectors from scratch," *IEEE International Conference on Computer Vision (ICCV)*, Venice, 2017, pp. 1937-1945.
- [100] M. R. Islam, D. Massicotte, P. Massicotte and W-P Zhu, "Domain Adaptation with Low-Latency Shallow Convolutional Neural Network for Improved Inter-Session/Inter-Subject Gesture Recognition" in submission to *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 2023.
- [101] Y. Wu and K. He, "Group normalization," In *ECCV*, pp. 3-19, 2018.
- [102] X. Li, S. Chen, X. Hu and J. Yang, "Understanding the Disharmony Between Dropout and Batch Normalization by Variance Shift," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 2019, pp. 2677-2685, doi: 10.1109/CVPR.2019.00279.
- [103] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," [Online]. Available: In arxiv:cs/arXiv:1409.1556, 2014.
- [104] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," [Online]. Available: arXiv preprint arXiv:1505.00853, 2015.
- [105] A. H. Al-Timemy, G. Bugmann, J. Escudero and N. Outram, " A preliminary investigation of the effect of force variation for myoelectric control of hand prosthesis," *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Osaka, Japan, 2013, pp. 5758-5761, doi: 10.1109/EMBC.2013.6610859.
- [106] F.J. Huang and Y. LeCun, "Large-scale Learning with SVM and Convolutional Nets for Generic Object Categorization," *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 284-291, 2006.
- [107] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, vol. 2, pp. 886–893, 2005.
- [108] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, Sept. 2010.

- [109] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 3, pp. 313–323, May 2012.
- [110] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [111] N. Dalal, "Finding people in images and videos," Ph.D. thesis, INRIA Rhone-Alpes, 2006.
- [112] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.
- [113] C.W. Hsu and C.J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 13, pp. 415–425, 2002.
- [114] Bristow, Hilton and Lucey, Simon, "Why do linear SVMs trained on HOG features perform so well?" in arXiv preprint arXiv:1406.2419, 2014.
- [115] C.-C. Chang and C.-J. Lin., "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, 2:27:1—27:27, 2011.
- [116] C. Hsu, C. C. Chang, and C. J. Lin., "A practical guide to support vector classification," Technical report, 2005.
- [117] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The binormal assumption on precision-recall curves," in Proc. 20th ICPR, Aug. 2010, pp. 4263–4266.
- [118] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding," in *ICLR* 2016.
- [119] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le, "Attention augmented convolutional networks," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 3285–3294.
- [120] B. Sun and K. Saenko., "Deep coral: Correlation alignment for deep domain adaptation," In *Computer Vision–ECCV 2016 Workshops*, pages 443–450. Springer, 2016
- [121] E. Tzeng, J. Hoffman, K. Saenko and T. Darrell, "Adversarial Discriminative Domain Adaptation," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 2962–2971, doi: 10.1109/CVPR.2017.316.
- [122] Ming-Yu Liu and Oncel Tuzel, "Coupled generative adversarial networks,". In *NIPS*, pages 469–477. 2016
- [123] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer., "Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 mb model size," arXiv preprint arXiv:1602.07360, 2016
- [124] R. N. Khushaba, "Correlation Analysis of Electromyogram Signals for Multiuser Myoelectric Interfaces," in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 4, pp. 745–755, July 2014, doi: 10.1109/TNSRE.2014.2304470.
- [125] Accessed online, Retrieved from (2018): [Kinova® Gen3 Ultra lightweight robot User Guide \(kinovarobotics.com\)](http://kinovarobotics.com)

## **Appendix A – Published Articles**

M. R. Islam, D. Massicotte, F. Nougrou, P. Massicotte and W. -P. Zhu, "S-Convnet: A Shallow Convolutional Neural Network Architecture for Neuromuscular Activity Recognition Using Instantaneous High-Density Surface EMG Images," 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Montreal, QC, Canada, 2020, pp. 744-749, doi: 10.1109/EMBC44109.2020.9175266.

# S-CONVNET: A SHALLOW CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE FOR NEUROMUSCULAR ACTIVITY RECOGNITION USING INSTANTANEOUS HIGH-DENSITY SURFACE EMG IMAGES

Md. Rabiul Islam, Daniel Massicotte, Francois Nougrou, Philippe Massicotte, and Wei-Ping Zhu\*

Dept. of Electrical and Computer Engineering, Université du Québec à Trois-Rivières, QC, Canada

\*Dept. of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada

**ABSTRACT**—The recent progress in recognizing low-resolution instantaneous high-density surface electromyography (HD-sEMG) images opens up new avenues for the development of more fluid and natural muscle-computer interfaces. However, the existing approaches employed a very large deep convolutional neural network (ConvNet) architecture and complex training schemes for HD-sEMG image recognition, which requires learning of >5.63 million (M) training parameters only during *fine-tuning* and *pre-trained* on a very large-scale labeled HD-sEMG training dataset, as a result, it makes high-end resource-bounded and computationally expensive. To overcome this problem, we propose S-ConvNet models, a simple yet efficient framework for learning instantaneous HD-sEMG images from scratch using *random-initialization*. Without using any pre-trained models, our proposed S-ConvNet demonstrate very competitive recognition accuracy to the more complex state of the art, while reducing learning parameters to only  $\approx 2M$  and using  $\approx 12 \times$  smaller dataset. The experimental results proved that the proposed S-ConvNet is highly effective for learning discriminative features for instantaneous HD-sEMG image recognition, especially in the data and high-end resource-constrained scenarios.

**Keywords:** Neuromuscular activity recognition, Shallow convolutional neural networks, Feature learning, HD-sEMG, Gesture recognition, Muscle-computer interface, Deep neural networks.

## I. INTRODUCTION

Neuromuscular activity recognition has a growing motivation for research because of its respective novel applications in real life. The major application domains are non-invasive control of active prosthesis [1], wheelchairs [2], exoskeletons [3] or providing interaction methods for video games [4] and neuromuscular diagnosis [5]. The conventional approaches for neuromuscular activity recognition immensely rely on sparse multi-channel surface electromyography (sEMG) sensors and windowed descriptive and discriminatory sEMG features [6-10]. However, the sparse multi-channel sEMG based methods are not suitable for real-world applications due to their lack of robustness to electrode shift and positioning and therefore malfunctioning in any one of the channels requires retraining the entire system [11], [12]. In recent years, the high-density sEMG (HD-sEMG) based methods have been proposed to address this problem [11-13], [29]. The HD-sEMG consists of two-dimensional (2D) arrays of closely spaced electrodes that used to record the myoelectric activity over the skin surface [13], [14].

The recorded HD-sEMG data are spatially correlated enabled both spatial and temporal changes and robust to electrode shift and positioning [12]. The windowed sEMG and descriptive and discriminative features are used by the existing HD-sEMG based methods for neuromuscular activity recognition. However, finding an optimal window size would still require that reflects the compromise between classification accuracy and controller delay (both increase with the window increase) especially in the application of assistive technology, physical rehabilitation, and human-computer interfaces [13].

To address this problem, the distinctive patterns inside the instantaneous sEMG images has been explored for developing more fluid and natural muscle-computer interfaces (MCI's) in recent years by Geng *et al.*, [13] and M. R. Islam *et al.*, [15], [33]. This scheme enables neuromuscular activity recognition solely with the sEMG images spatially composed from HD-sEMG signals recorded at a specific instant. The instantaneous values of HD-sEMG signals at each sampling instant were arranged in a 2D grid following the electrode positioning. Afterwards, this 2D grid was converted to a grayscale sEMG image. Using Histogram of Oriented Gradients (HOG) as discriminative features and pairwise SVM's classification method in [15], a competitive neuromuscular activity recognition accuracy of an 8-hand gesture has been achieved as par with the state-of-the-art method for an intra-subject test.

However, a DeepFace [17] like very large deep convolutional neural network (CNN or ConvNet) architecture is employed by the state-of-the-art methods [13], [16] for sEMG image classification, which requires to be pre-trained on a very large-scale training dataset ( $\approx 0.76$  million), as a result, it makes computationally expensive to be practical for real-world MCIs applications. Following are the other critical limitations of using *pre-trained* networks for instantaneous HD-sEMG image recognition:

(i) *Constrained structure design space* – pre-trained networks are very deep and large and trained on a large-scale HD-sEMG dataset, therefore, containing a massive number of parameters. Hence, there is a little flexibility to control/adjust the network structures (even for small changes) by directly adopting the *pre-trained* network to the *target task*. The requirement of computing resources and large-scale pre-trained datasets are also bounded by large network structures.

(ii) *Domain mismatch* – the existing sEMG based neuromuscular activity recognition methods are usually trained and evaluated on the data acquired from the able-bodied subjects. However, in real time

sEMG-based MCIs applications (e.g., assistive technology, physical rehabilitation, etc.) are most of the time designed for elderly people, amputees and patients. These differences impose a serious problem due to the varied sEMG distributions in the *source* and *target task*. Though the fine-tuning of the pre-trained model can reduce the gap, however, it is still a serious problem, when there is a huge mismatch between the *source* and the *target task* [18]. Also, this conventional wisdom of pre-training is recently challenged by He *et al.* [31], where *pre-training* does not necessarily improve the *target task* accuracy is proved to be claimed.

(iii) *Learning bias* – the distributions and the loss functions between the *source task* and the *target task* may vary significantly, which may lead to different searching/optimization spaces. Therefore, the learning may be biased towards a local minimum which is not optimal for the *target task* [19].

To overcome these above-mentioned problems, our work is motivated by the following research question- *is it possible to learn neuromuscular activities from scratch utilizing HD-sEMG datasets available only for the target task without any pre-training?* To achieve this goal, we propose shallow and lightweight convolutional neural network (S-ConvNet) architectures, a simple yet effective framework, which could learn neuromuscular activity from scratch using  $\approx 12 \times$  smaller dataset than its pre-trained counterparts for HD-sEMG image recognition.

For instantaneous sEMG image-based neuromuscular activity recognition, the challenge remains open because very limited research has been done on it.

The rest of the paper is organized as follows. Sections II and III present the proposed framework and S-ConvNet models respectively. Section IV describes the testing database and experimental validation. Section V offers some conclusive remarks.

## II. THE FRAMEWORK

The proposed S-ConvNet framework for neuromuscular activity recognition using instantaneous HD-sEMG images has three phases: (i) pre-processing and HD-sEMG image generation (ii) architectural design of the S-ConvNet model and (iii) classification. Fig.1 describes the proposed S-ConvNet framework of muscular activity recognition by instantaneous sEMG images. The All-ConvNet [33] is a fully convolutional neural network, where the depth of the All-ConvNet network is much higher than the proposed S-ConvNet. First, the power-line interferences were removed from the acquired HD-sEMG signals with a band-stop filtered between 45 and 55 Hz using a 2<sup>nd</sup> order Butterworth filter. Then, the HD-sEMG signals at each sampling instant were arranged in a 2-D grid according to their

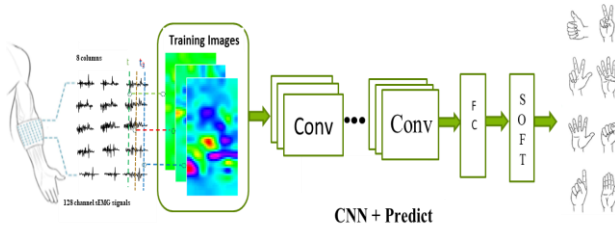


Fig. 1. Schematic diagram of the proposed framework of muscular activity recognition by instantaneous sEMG images.

electrode positioning. This grid was further transformed into an instantaneous sEMG image by linearly transforming the values of sEMG signals from  $mV$  to color intensity as  $[-2.5mV, 2.5mV]$  to  $[0 255]$ . Thus, an instantaneous grayscale sEMG image was formed with a size of  $16 \times 8$ . Secondly, we devised different S-ConvNet models which describe in Section III. Finally, providing instantaneous HD-sEMG images and their corresponding labels, our devised S-ConvNet model is trained offline to predict to which muscular activity an instantaneous HD-sEMG image belongs. Then, this trained S-ConvNet model is used to recognize different neuromuscular activities at test time from the unseen instantaneous HD-sEMG images.

## III. MODEL DESCRIPTION- THE SHALLOW CONVOLUTIONAL NEURAL NETWORK (S-CONVNET)

We train our S-ConvNet on a multi-class neuromuscular activity recognition task, namely, to recognize an activity class through an instantaneous HD-sEMG image. The overall architecture of S-ConvNet models are described in Table I. Starting from the simplest Model A, the depth and number of parameters in the network gradually increase to Model C. The instantaneous HD-sEMG image is passed through a convolutional (conv.) layers, where a small receptive field with a  $3 \times 3$  filters are used. The smallest receptive field with  $3 \times 3$  filters is the minimum filter size to allow overlapping convolutions and spatial pooling with a stride of 2, which also captures the notion of left, right and center amicably. It can be observed that the Model B from the Table I is a variant of the Network in Network architecture [24], where only  $1 \times 1$  convolution is performed after each normal  $3 \times 3$  convolutions layers. The  $1 \times 1$  convolution act as a linear transformation of the input channels followed by a non-linearity [25]. We also highlight that the model C is a variant of the simple ConvNet models introduced by J. T. Springenberg *et al.*, [20] for object recognition in which the spatial pooling is performed by using a stridden CNN. The output of a convolution map  $f$  produced by a convolution layer  $c$  is computed as follows:

$$c_{i,j,o}(f) = \phi \left( \sum_{h=1}^k \sum_{w=1}^k \sum_{u=1}^n \theta_{h,w,u,o} \cdot f_{g(h,w,i,j,u)} \right) \quad (1)$$

where  $\theta$  are the convolutional weights or filters;  $g(h,w,i,j,u) = (r \cdot i + h, r \cdot j + w, u)$  is the function mapping from a position in  $c$  to a position in  $f$  respecting the stride  $r$ ;  $w$  and  $h$  are respectively the width and height of the filters;  $n$  is the number of channels (in case  $f$  is the output of a convolutional layer,  $n$  is the number of filters);  $o \in [1, M]$  is the number of output feature or channels of the convolutional layer and  $\phi(\cdot)$  is the activation function, an exponential linear unit ELU defined as:

$$\phi(x) = \begin{cases} \alpha(\exp(x) - 1), & \text{if } x < 0 \\ x, & \text{if } x \geq 0 \end{cases} \quad (2)$$

Afterwards, the convolution maps produced by the final convolutional layer of each of the model networks, illustrated in Table I, are flattened out to form a multi-dimensional feature vector. Then, the flattened feature vector is inputted to a fully connected layer where each of the feature elements is connected to all its input neurons. This fully connected layer can capture correlations between features extracted in the distant part of the instantaneous sEMG images.

TABLE I THE THREE S-CONVNET NETWORKS MODELS FOR NEUROMUSCULAR ACTIVITY RECOGNITION

A	B	C
<b>Input 16×8 Gray-level Image</b>		
3 × 3 Conv. 32 ELU	3 × 3 Conv. 32 ELU	3 × 3 Conv. 32 ELU
3 × 3 Conv. 64 ELU	1 × 1 Conv. 32 ELU	3 × 3 Conv. 32 ELU
3 × 3 Conv. 64 ELU	3 × 3 Conv. 64 ELU	3 × 3 Conv. 32 ELU, with stride $r = 2$
FC1 256 ELU	1 × 1 Conv. 64 ELU	3 × 3 Conv. 64 ELU
FC2 G-way softmax	FC1 256 ELU	3 × 3 Conv. 64 ELU
-	FC2 G-way softmax	3 × 3 Conv. 64 ELU, with stride $r = 2$
-	-	FC1 256 ELU
-	-	FC2 G-way softmax

Finally, the output of the fully connected layer is fed to a  $G$ -way softmax layer (where  $G$  is the number of neuromuscular activity classes) which produces a distribution over the class labels. If we de-

note  $\hat{y}^{(j)}$  as the  $j$ th element of the  $G$  dimensional output vector of the layer preceding the softmax layer, the class probabilities are estimated using the softmax function  $\sigma(\cdot)$  defined as below:

$$\sigma(\hat{y}^{(j)}) = \frac{\exp(\hat{y}^{(j)})}{\sum_G \exp(\hat{y}^{(G)})} \quad (3)$$

The goal of this training is to maximize the probability of the correct neuromuscular activity class. We achieve this by minimizing the cross-entropy loss [26] for each training sample. If  $y$  is the true label for a given input, the loss is

$$L = -\sum_j y^{(j)} \ln(\sigma(\hat{y}^{(j)})) \quad (4)$$

The loss is minimized over the parameters by computing the gradient of  $L$  with respect to the parameters and by updating the parameters using the state-of-the-art Adam (adaptive moment estimation) gradient descent-based optimization algorithm [27].

Having trained the network, an instantaneous HD-sEMG image is recognized as in the neuromuscular activity class  $C$  by simply propagating the input image forward and computing:  $C = \text{argmax}_j(\hat{y}^{(j)})$ .

The major advantage of the proposed S-ConvNet models are easily scalable and does not increase the learning parameters with the enhancement of input HD-sEMG image size. Whereas, the ConvNet employed by the state of the art [13] is unscalable. For example, the learning parameters of [13] increase to  $\approx 5.63\text{M}$  to  $\approx 11\text{M}$  with a little augmentation of input HD-sEMG image size from  $16 \times 8$  to  $16 \times 16$  due to the use of an unshared weight strategy [32].

#### IV. THE PERFORMANCE EVALUATION OF THE PROPOSED S-CONVNET MODELS

In order to quantify the effect of simplifying the proposed S-ConvNet model architecture, we perform experiments on CapgMyo data sets [16] (These data sets are made publicly available from the following website: <http://zju-capg.org/myo/data/index.html>). This dataset was developed for providing a standard benchmark database (DB) to explore new possibilities for studying next-generation muscle-computer interfaces (MCIs). The CapgMyo database comprises 3 sub-databases (referred as DB-a, DB-b and DB-c). However, DB-a has been used in our experiments to evaluate the

TABLE II THE AVERAGE RECOGNITION ACCURACY (%) OF 8 HAND GESTURES WITH INSTANTANEOUS HD-SEMG IMAGES FOR 18 DIFFERENT SUBJECTS AND RECOGNITION APPROACHES. MAJORITY VOTING (ON 40 SEMG IMAGE) RESULTS ARE SHOWN IN PARENTHESES

Model	Average Recognition Accuracy (%)	# Learning Parameters
<b>S-ConvNet-A</b>	<b>87.95 (98.87)</b>	$\approx 2.09\text{M}$
S-ConvNet-B	86.94	$\approx 2.12\text{M}$
S-ConvNet-C	87.02	$\approx 2.10\text{M}$
W.Geng <i>et al.</i> , [13]	89.3 (99.00)	$\approx 5.63\text{M} + \text{Pre-training}$

performance of our proposed methods for *intra-session* neuromuscular activity recognition because the maximum number of subjects (18) have participated in DB-a. In DB-a, 8 isotonic and isometric hand gestures were obtained from 18 of the 23 subjects and

each gesture was also recorded 10 times. For each subject, the recorded HD-sEMG data is filtered, sampled and instantaneous sEMG image is generated using the method described in Section II. More explicitly, 8 different hand gestures are performed by every subject and each hand gestures are recorded 10 times with a 1000 Hz sampling rate, which in total generates  $(8 \times 10 \times 1000) = 80\,000$  or 80k instantaneous sEMG images individually. Then, our S-ConvNet models are learned from *scratch* through *random initialization*. We performed training, validation and testing using only 80 000 images produced by 18 subjects individually through a leave one trial out cross-validation. We kept one trial out from each of the 8 different hand gestures i.e 8 000 images for validation and testing. The remaining 9 trials for 8 different hand gestures i.e 72k images have been used for training. The cross-validation accuracy  $A$  is computed for each class  $i$ , as the number of totals correctly recognized hand gestures, divided by the total number of tests sEMG images

$$\text{Accuracy, } A = \frac{C}{N} = \frac{\sum C_i}{N} \quad (5)$$

where  $i = \{1, 2, \dots, G\}$  and  $G$  is the number of gesture classes.

In contrast, existing approaches (e.g., [13] and [16]) for instantaneous HD-sEMG image recognition used a total of  $(18 \times 40\,000) = 720\,000$  or 720k training images for pre-training, while 40 000 images from each of the subject are used separately for *fine-tuning*. Therefore, the existing approaches involve a total of  $(720\,000 + 40\,000) = 760\,000$  or 760k images only in the training process.

In our experiments, we compared all the proposed S-ConvNet models described in Section III on the CapgMyo DB-a datasets without any *pre-training* or data augmentations. The connection weights for all S-ConvNet networks were *randomly initialized* using Xavier and He initialization schemes [21], [28]. However, we found that the models with He initialization scheme perform on average 1-1.5% worse than the Xavier initialization. We also propose to use a computationally efficient stochastic optimization algorithm, Adam [27], which provides fast and reliable learning convergence than the stochastic gradient descent (SGD) optimization algorithm used in the literature for instantaneous HD-sEMG image recognition. Our proposed all S-ConvNet models were trained using Adam optimization algorithms with a momentum decay and scaling decay are initialized to 0.9 and 0.999 respectively. In contrast to SGD, Adam is an adaptive learning rate algorithm, therefore, it requires less tuning of the learning rate hyperparameter. The learning rate of 0.001 is initialized to all our experiments. The smaller batches of 256

randomly chosen samples from the training dataset are fed to the network during consecutive learning iterations for all our experiments. We set a maximum of 100 epochs for training our S-ConvNet models. However, to avoid overfitting we have also applied early stopping in which the training process is interrupted if no improvements in validation loss are noticed for 5 consecutive epochs. The Batch normalization [22] is applied after the input and before each non-linearity. The Dropout [23] was applied on all layers with probabilities 35% for all S-ConvNet models. The S-ConvNet models were trained on a workstation with an Intel Core, 3.60 (i7-4790) processor, 16GB RAM and an NVIDIA RTX Ti GPU. Each epoch was completed in approximately 4s while training with S-ConvNet-A. The test results for all the S-ConvNet models are presented in Table II and compared with state-of-the-art methods.

As can be seen in the Table II, the simple S-ConvNet models (on the order of  $\approx 2M$  learning parameters) trained from *random-initialization* with  $3 \times 3$  convolutions and a dense layer with only a smaller number of neuron performs comparably to the state of the art for CapgMyo DB-a dataset even though the state of the art methods use more complicated network architectures and training schemes which requires to learn over  $\approx 5.63M$  parameters during fine-tuning only and also pre-trained with over 720k instantaneous HD-sEMG images.

Fig. 2 presents the recognition accuracy obtained by our proposed different S-ConvNet models for 18 different subjects and their statistical significance. We achieve 87.95%, 86.94%, and 87.02% average recognition accuracy for the proposed S-ConvNet-A, B, and C models respectively, which is very competitive to the more complex, highly resource-based and fine-tuned pre-trained models proposed by the existing approaches while also reducing the learning parameters to a large extent. These high recognition accuracies for neuromuscular activity recognition based on instantaneous HD-sEMG images indicate the stability and potentiality of the proposed S-ConvNet models.

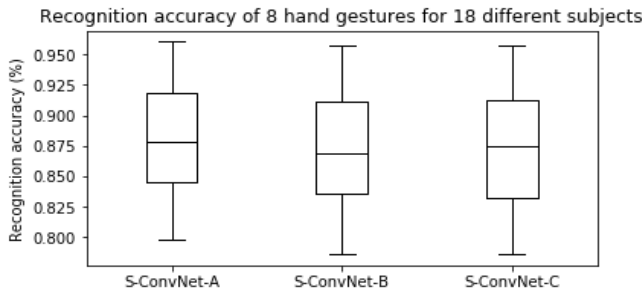


Fig. 2. The recognition accuracy of 8 hand gestures for 18 different subjects with our proposed S-ConvNet recognition approaches.

The recognition accuracy of 8 hand gestures of all 18 subjects in CapgMyo DB-a which obtained through leave one trial out cross-validation for 10 different trials using S-ConvNet-A and their statistical significance are presented in Fig. 3. It is observed that the average recognition accuracy  $>93\%$  and  $>88\%$  have been achieved at least for 6 and 5 different subjects respectively. Moreover, the high average recognition accuracies 94.29%, 96.55% and 98.87% are achieved by a simple majority voting with 3, 20 and 40 instantaneous images respectively (Table II, S-ConvNet-A). All these highly promising and competitive results proved that the proposed S-

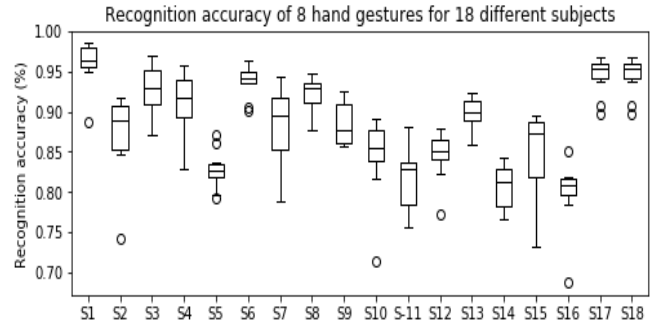


Fig. 3. The recognition accuracy of 8 hand gestures for 18 different subjects with our proposed S-ConvNet-A.

ConvNet models trained from *random-initialization* can learn all the necessary invariances that require to build a discriminant representation using only the available target dataset for neuromuscular activity recognition based on instantaneous HD-sEMG images. Therefore, the performance of the proposed S-ConvNet models is no worse than its more complex, highly resource-based, pre-trained and fine-tuned state of the art models. It is also worth mentioning that why we did compare our results only with [13] and not with the [16] and [30]. Because the same pre-trained and fine-tuned model employed in [13] was used in these successive studies, however, to address the same problem with a different view. Now, for the fair comparison with the state of the art, the following points are required to be highlighted.

We introduce a leave one trial out cross-validation in which our proposed S-ConvNet models are tested with 80k different samples for every subject. Existing instantaneous HD-sEMG image recognition approaches are tested with 40k samples for each of the subjects. Whereas we have used 80k samples (twice the number of testing samples) for recognition and achieved comparable performance on par with the state of the art. It is also noteworthy that the recognition results of all S-ConvNet models are obtained without any hyper-parameter tuning. Therefore, we also want to stress out that the results of all models evaluated in this paper could potentially be improved or even surpass the state of the art by a thorough hyperparameter tuning.

Finally, we argue that as aforementioned briefly, training from scratch is of critical importance at least for the following reasons. First, *Domain mismatch*—the distributions of the sEMG signals vary considerably even between recording sessions of the same subject. This problem becomes even more challenging, where the learned model is used to recognize muscular activities in a different recording session. Though the fine-tuning of the pre-trained model can reduce the gap due to the deformations in a new recording session. But what if we have a technique that can learn HD-sEMG images from scratch for recognizing neuromuscular activities. Second, the fine-tuned pre-trained model restricts the structure design space for neuromuscular activity recognition. This is very critical for the deployment of deep neural network models to the resource limited scenarios.

## V. CONCLUSION

We present S-ConvNet models, a simple yet efficient framework for learning instantaneous HD-sEMG images from



scratch for neuromuscular activity recognition. Without using any pre-trained models, our proposed S-ConvNet demonstrates very competitive accuracy to the more complex state of the art for neuromuscular activity recognition based on instantaneous HD-sEMG images, while using  $\approx 12\times$  smaller dataset and reducing learning parameters to  $\approx 2M$ . The proposed S-ConvNet has great potential for learning and recognizing neuromuscular activities on resource-bounded devices. Our future work will consider improving inter-session neuromuscular activity recognition performances, as well as learning S-ConvNet models to support resource-bounded devices.

#### ACKNOWLEDGMENT

This work was supported in part by the Regroupement stratégique en microsystèmes du Québec (ReSMiQ) and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

#### REFERENCES

- [1] D. Farina *et al.*, "The extraction of neural information from the surface EMG for the control of upper-limb prostheses: merging avenues and challenges," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, Vol 22, no. 2, pp. 797–809, 2014.
- [2] G. Jang, J. Kim, J. S. Lee, and Y. Choi, "EMG-based continuous control scheme with simple classifier for electric-powered wheelchair," *IEEE Transactions on Industrial Electronics*, Vol. 63, no. 6, pp. 3695–3705, 2016.
- [3] R. Jimenez-Fabian, and O. Verlinden, "Review of control algorithms for robotic ankle systems in lower-limb orthoses, prostheses, and exoskeletons," *Medical Engineering & Physics*, Vol 34, no. 4, pp. 397–408, 2012.
- [4] D.-H. Kim, *et al.*, "Epidermal electronics," *Science* 333, 838–843, 2011.
- [5] Y. Hu, J. N. Mak, & K. Luk, "Application of surface EMG topography in low back pain rehabilitation assessment," *International IEEE/EMBS Conference on Neural Engineering*, pp. 557–560, 2007.
- [6] E. Costanza, S. A. Inverso, R. Allen, R. and P. Maes, "Intimate interfaces in action: assessing the usability and subtlety of EMG-based motionless gestures," *Conference on Human Factors in Computing Systems*, ACM, pp. 819–828, 2007.
- [7] T. S. Saponas, D. S. Tan, D. Morris, and R. Balakrishnan, "Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces," *Conference on Human Factors in Computing Systems*, ACM, pp. 515–524, 2008.
- [8] T. S. Saponas, D. S. Tan, D. Morris, D. J. Turner and J. A. Landay, "Making muscle-computer interfaces more practical," *Conference on Human Factors in Computing Systems*, pp. 851–854, ACM, 2010.
- [9] M. Atzori *et al.*, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Scientific Data* 1, 2014.
- [10] N. Patricia, T. Tommasi, & B. Caputo, "Multi-source adaptive learning for fast control of prosthetics hand," *International Conference on Pattern Recognition*, pp. 2769–2774, 2014.
- [11] C. Amma, T. Krings, J. Ber, J. and T. Schultz, "Advancing muscle-computer interfaces with high-density electromyography," *Conference on Human Factors in Computing Systems*, pp. 929–938, ACM, 2015.
- [12] A. Stango, F. Negro and D. Farina, "Spatial correlation of high density EMG signals provides features robust to electrode number and shift in pattern recognition for myocontrol," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* Vol 23, no.2, pp. 189–198, 2015.
- [13] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu and J. Li, "Gesture recognition by instantaneous surface EMG images," *Sci. Rep.*, Vol 15, no.6, 36571, Nov 2016.
- [14] R. Casale and A. Rainoldi, "Fatigue and fibromyalgia syndrome: clinical and neurophysiologic pattern," *Best Practice & Research Clinical Rheumatology* Vol 25, no. 2, 241–247, 2011.
- [15] M. R. Islam, D. Massicotte, F. Nougrou and W. Zhu, "HOG and Pairwise SVMs for Neuromuscular Activity Recognition Using Instantaneous HD-sEMG Images," *2018 16th IEEE International New Circuits and Systems Conference (NEWCAS)*, Montreal, QC, 2018, pp. 335-339.
- [16] Y. Du., W. Jin, W. Wei, Y. Hu and W Geng, "Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation," *Sensors*, Vol. 17, no. 3, 458, 2017.
- [17] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: Closing the Gap to Human-Level Performance in Face Verification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, pp. 1701-1708, 23-28 June, 2014.
- [18] S. Gupta, J. Hoffman, and J. Malik., "Cross modal distillation for supervision transfer," *In CVPR*, 2016.
- [19] Z. Shen, Z. Liu, J. Li, Y.-G. Jiang, Y. Chen, and X. Xue., "DSOD: Learning deeply supervised object detectors from scratch," *In ICCV*, 2017.
- [20] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller., "Striving for Simplicity: The All Convolutional Net," *CoRR*, abs/1412.6806, 2014.
- [21] X. Glorot and Y. Bengio., "Understanding the difficulty of training deep feedforward neural networks," *In AISTATS*, 2010.
- [22] S. Ioffe and C. Szegedy., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *In ICML*, 2015.
- [23] Srivastava, Nitish, Hinton, Geoffrey E, Krizhevsky, Alex, Sutskever, Ilya, and Salakhutdinov, Ruslan., "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, Vol 15, no. 1, pp.1929–1958, 2014.
- [24] Lin, Min, Chen, Qiang, and Yan, Shuicheng., "Network in network," *In ICLR: Conference Track*, 2014.
- [25] Simonyan, Karen and Zisserman, Andrew., "Very deep convolutional networks for large-scale image recognition," *In arxiv:cs/arXiv:1409.1556*, 2014.
- [26] Janocha, Katarzyna, and Wojciech Marian Czarnecki. "On loss functions for deep neural networks in classification." *arXiv preprint arXiv:1702.05659*, 2017.
- [27] Kingma, Diederik and Ba, Jimmy., "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [28] K. He, X. Zhang, S. Ren, and J. Sun., "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," *In ICCV*, 2015.
- [29] F. Nougrou, A. Campeau-Lecours, R. Islam, D. Massicotte and B. Gosselin, "Muscle Activity Distribution Features Extracted from HD sEMG to Perform Forearm Pattern Recognition," *2018 IEEE Life Sciences Conference (LSC)*, Montreal, QC, pp. 275-278, Oct. 2018.
- [30] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli and W. Geng, "Surface-Electromyography-Based Gesture Recognition by Multi-View Deep Learning," in *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2964-2973, Oct. 2019.
- [31] K. He, R. Girshick, and P. Dollar, "Rethinking imagenet pretraining", *In ICCV*, 2019.
- [32] M. R. Islam, D. Massicotte, F. Nougrou, P. Massicotte and W. Zhu., "S-ConvNet: A shallow convolutional neural network architecture for neuromuscular activity recognition using instantaneous high-density surface EMG images," *arXiv preprint arXiv:1906.03381*, 2019.
- [33] M. R. Islam, D. Massicotte and W. Zhu., "All-ConvNet: A Lightweight All CNN for Neuromuscular Activity Recognition Using Instantaneous High-Density sEMG Images," *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Dubrovnik, Croatia, 2020, pp. 1-5.

M. R. Islam, D. Massicotte and W. -P. Zhu, "All-ConvNet: A Lightweight All CNN for Neuromuscular Activity Recognition Using Instantaneous High-Density Surface EMG Images," *2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Dubrovnik, Croatia, 2020, pp. 1-6, doi: 10.1109/I2MTC43012.2020.9129362.

# All-ConvNet: A Lightweight All CNN for Neuromuscular Activity Recognition Using Instantaneous High-Density Surface EMG Images

Md. Rabiul Islam, Daniel Massicotte, and Wei-Ping Zhu\*

Dept. of Electrical and Computer Engineering, Université du Québec à Trois-Rivières, QC, Canada

\*Dept. of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada

{md.rabiul.islam, daniel.massicotte}@uqtr.ca, weiping@ece.concordia.ca

**ABSTRACT** – Neuromuscular activity recognition using low-resolution instantaneous high-density surface electromyography (HD-sEMG) images present a great challenge. The recent result shows the high potentiality and hence opens up new avenues for the development of more fluid and natural muscle-computer interfaces. However, the existing approaches employed a very large deep ConvNet, which requires learning >5.63 million training parameters only during *fine-tuning* and *pre-trained* on a very large-scale labeled HD-sEMG training datasets, as a result, it makes high-end resource bounded and computationally expensive. To overcome this problem, we propose a lightweight All-ConvNet model that consists solely of convolutional layers, a simple yet efficient framework for learning instantaneous HD-sEMG images from scratch through *random initialization*. Without using any pre-trained models, our proposed lightweight All-ConvNet demonstrate very competitive or even state of the art performance on a current benchmarks HD-sEMG dataset, while requires learning only  $\approx 460k$  training parameters and using  $\approx 12\times$  smaller dataset. The experimental results proved that the proposed lightweight All-ConvNet is highly effective for learning discriminative features for low-resolution instantaneous HD-sEMG image recognition and low-latency processing especially in the data and high-end resource constrained scenarios.

**Keywords:** Neuromuscular activity recognition, All convolutional neural networks, Feature learning, HD-sEMG, Gesture recognition, Muscle-computer interface, Deep neural networks.

## I. INTRODUCTION

Neuromuscular activity recognition has been a driving motivation for some emerging biomedical applications such as non-invasive and intuitive control of active prosthesis, wheelchairs, exoskeletons or providing interaction methods for video games and neuromuscular diagnosis [1-4]. The sparse-channel surface electromyography (sEMG) and windowed descriptive and discriminative sEMG features are used by the conventional approaches [5-9], [28]. However, these methods are not practical for high sensitivity to electrode shift and positioning [10-11]. To overcome this problem, the high-density sEMG (HD-sEMG) based methods have been proposed in recent years [10-14], [26-27]. The HD-sEMG records myoelectric signals using two-dimensional (2D) electrode arrays that characterize the spatial distribution of myoelectric activity over the muscles that reside within the electrode pick-up area [12]. The collected HD-sEMG data are spatially correlated which enabled both temporal and spatial

changes and robust against malfunction of the channels with respect to the previous counterparts [11]. However, the existing HD-sEMG based neuromuscular activity recognition methods [26-27], [28] are still depending on the windowed sEMG (e.g., 260 ms) which demands to find an optimal window length otherwise influence in the classification accuracy and controller delay especially in the application of assistive technology, physical rehabilitation and human computer interfaces [12].

To overcome this problem and develop a more fluid and natural muscle-computer interfaces (MCI's), more recently, W. Geng *et al.*, [12] and M. R. Islam *et al.*, [13] explored the patterns inside the instantaneous sEMG images spatially composed from HD-sEMG enables neuromuscular based gesture recognition solely with the sEMG signals recorded at a specific instant. Hence, the observational latency was reduced to 1 ms which would reduce controller delay significantly to the above-mentioned applications.

However, the current state-of-the-art methods [12], [14] employed a DeepFace [15] like very large deep convolutional neural network (CNN or ConvNet) architecture for sEMG image classification, which requires learning >5.63 million (M) training parameters only during *fine-tuning* and *pre-trained* on a very large-scale labeled HD-sEMG training datasets, as a result it makes high-end resource bounded and computationally expensive to be practical for real-world MCIs applications. The major limitations of using pre-trained networks are usually very deep, contains a massive number of parameters and trained on a large-scale training dataset. Therefore, it is totally not possible to any degree of mutation of the pre-trained networks during fine-tuning. If any mutation or employing a new architecture is necessary then the whole pre-training should be re-conducted on the large-scale training dataset, requiring a high computational cost. Fortunately training from scratch can cope with these problems [29].

Moreover, in their pre-trained ConvNet includes two locally connected (LCN) and three fully connected layers among the other convolutions and a  $G$ -way fully connected layer. The LCN layers assign an independent filter weight,  $\theta_p$  to each of the local receptive field of a feature map i.e.  $f_p = I_p^T \theta_p$ ,<sup>1</sup> while convolution (or CNN) layers adopt a filter weight sharing

<sup>1</sup> Given an input sEMG image  $I$ , LCN requires each filter is conducted on a patch vector  $I_p$ , where  $p$  stands for position of the patch in the input image.

strategy i.e.  $f_p = I_p^T \theta$  [16]. Due to this unshared weight strategies of LCN, the number of learning parameters increases considerably from  $m$  to  $m \times k$ , where  $m \gg k$ , where  $m$  is the number of patches and  $k$  is the number of kernels. As a result, a very large-scale labeled training dataset is required to train the LCN [15]. However, the LCN may be useful in an application where the precise location of the feature is dependent of the class labels.

Considering the above-mentioned fact, we must investigate - (i) *Do we expect the devised networks model to produce a location/translation invariant feature representation?* or, (ii) *do we need a location dependent feature representation?* Following this finding and building on other recent works for finding a simple network architecture, we propose a lightweight All-ConvNet, a new architecture that consists solely of convolutional layers, a simple yet effective framework, which could learn neuromuscular activity from scratch and yields competitive or even state of the art performance using  $\approx 12 \times$  smaller dataset while reducing learning parameters from  $\approx 5.63M$  to only  $\approx 460k$  than its pre-trained counterparts for instantaneous HD-sEMG image recognition.

For instantaneous sEMG image based neuromuscular activity recognition, the challenge remains open because very limited research has been done on it. We propose a lightweight All-ConvNet, to the best of our knowledge, this is the first All-ConvNet framework to date for instantaneous HD-sEMG recognition.

## II. THE PROPOSED FRAMEWORK

The proposed framework for neuromuscular activity recognition using instantaneous HD-sEMG images includes the following three major computational components: (i) pre-processing and sEMG image generation (ii) architectural design of the All-ConvNet model and (iii) classification. A schematic diagram of the proposed framework of muscular activity recognition by instantaneous sEMG images is shown in Fig. 1. First, the power-line interferences were removed from the acquired HD-sEMG signals with a band-stop filtered between 45 and 55 Hz using a 2<sup>nd</sup> order Butterworth filter. Then, the HD-sEMG signals at each sampling instant were arranged in a 2-D grid according to their electrode positioning. This grid was further transformed into an instantaneous sEMG image by linearly transforming the values of sEMG signals from  $mV$  to color intensity as  $[-2.5mV, 2.5mV]$  to  $[0, 255]$ . Thus, an instantaneous grayscale sEMG image was formed with a size of  $16 \times 8$ . Secondly, we devised a lightweight All-ConvNet model which describes in Section III. Finally, providing instantaneous HD-sEMG images and their corresponding labels, our devised All-ConvNet model is trained offline to predict to which muscular activity an instantaneous HD-sEMG image belongs. Then, this trained All-ConvNet model is used to recognize different neuromuscular activities at test time from the unseen instantaneous HD-sEMG images.

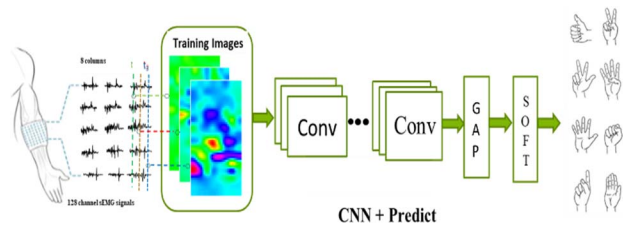


Fig. 1. Schematic diagram of the proposed framework of muscular activity recognition by instantaneous sEMG images.

## III. MODEL DESCRIPTION- THE ALL CONVOLUTIONAL NEURAL NETWORK (ALL-CONVNET)

The All-ConvNet architectural design is adopted based on the following principle and observation:

(i) It was hypothesized that the different muscular activities produce different spatial intensity distributions, which is reproducible across the trials of the same muscular activities and discriminative among different activities. However, we observed that the spatial intensity distributions within the same muscular activities are not locally invariant and the precise location of the features is also independent to the class labels. Fig. 2 demonstrate a sequence of HD-sEMG images derived from the same class. CNN alone has the great ability to exploit locally translational invariance features by adopting local connectivity and weight sharing strategies [16]. Hence, the LCN's are ablated in designing our All-ConvNet models as the location of the features is not dependent to the class labels. Why the ablation of LCN's are so significant? Because it is not only increased the number of training parameters but also make the network totally unscalable. For example, only the two LCN in [12] requires learning of  $> 2.13M$  parameters and the total learning parameters of [12] increased from  $\approx 5.63M$  to  $\approx 11M$  with just a little enhancement of input HD-sEMG image size from  $16 \times 8$  to  $16 \times 16$ .

(ii) Inspired by [17], we make use of the fact that if the part of the instantaneous HD-sEMG image is covered by the units in the topmost convolution layers could be large enough to recognize its content (i.e., muscular activity class we want to recognize). Then, the fully connected layers can also be replaced by simple 1-by-1 convolutions. This leads to predictions of HD-sEMG image classes at different positions which can then simply be averaged over the whole image. This scheme was first described by Lin *et. al.*, [21], which further regularizes the network as the 1-by-1 convolution has much less parameters than a fully connected and LCN layer. Overall

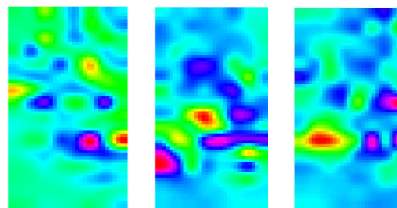


Fig. 2. HD-sEMGs derived from the same muscular activity class which demonstrates that the distributions are independent to the class labels.

TABLE I THE ALL-CONVNET NETWORK MODEL FOR NEUROMUSCULAR ACTIVITY RECOGNITION.

All-ConvNet
Input 16×16 Gray-level Image
3 × 3 Conv. 64 ELU
3 × 3 Conv. 64 ELU
3 × 3 Conv. 64 ELU with stride $r=2$
3 × 3 Conv. 128 ELU
3 × 3 Conv. 128 ELU
3 × 3 Conv. 128 ELU with stride $r=2$
1×1 Conv. 128 ELU
1×1 Conv. 8 ELU
global averaging over 4×4 spatial dimensions
G-way SoftMax

our architecture is thus reduced to consist only of convolutional layers with ELU non-linearities [25] and a global average pooling (GAP) + SoftMax layer to produce predictions over the whole instantaneous HD-sEMG image. Table I describes our proposed All-ConvNet architecture. We did experiments with the variant of All-ConvNet as in [17], however, the All-ConvNet presented in Table I performs favorably.

We train our All-ConvNet on a multi-class neuromuscular activity recognition task, namely, to recognize an activity class through an instantaneous HD-sEMG image. As described in the Table I, in the proposed All-ConvNet network we consider use 1-by-1 convolution at the top to produce 8 outputs of which we then compute an average over all positions and fed to a  $G$ -way SoftMax layer (where  $G$  is the number of neuromuscular activity classes) which produces a distribution over the class labels. If we denote  $\hat{y}^{(j)}$  as the  $j$ th element of the  $G$  dimensional output vector of the layer preceding the SoftMax layer, the class probabilities are estimated using the SoftMax function  $\sigma(\cdot)$  defined as below:

$$\sigma(\hat{y}^{(j)}) = \frac{\exp(\hat{y}^{(j)})}{\sum_G \exp(\hat{y}^{(e)})} \quad (1)$$

The goal of this training is to maximize the probability of the correct neuromuscular activity class. We achieve this by minimizing the cross-entropy loss [22] for each training sample. If  $y$  is the true label for a given input, the loss is

$$L = -\sum_j y^{(j)} \ln(\sigma(\hat{y}^{(j)})) \quad (2)$$

The loss is minimized over the parameters by computing the gradient of  $L$  with respect to the parameters and by updating the parameters using the state-of-the-art Adam (adaptive moment estimation) gradient descent-based optimization algorithm [23], which provides fast and reliable learning convergence than the stochastic gradient descent (SGD) optimization algorithm used in the fine-tuned pre-trained networks for instantaneous HD-sEMG image recognition.

Having trained the network, an instantaneous HD-sEMG image is recognized as in the neuromuscular activity class  $C$  by simply propagating the input image forward and computing:

$$C = \operatorname{argmax}_j(\hat{y}^{(j)}) \quad (3)$$

#### IV. THE PERFORMANCE EVALUATION OF THE PROPOSED ALL-CONVNET MODEL

In order to quantify the effect of simplifying the proposed All-ConvNet network we perform experiments on CapgMyo data sets [14] (These data sets are made publicly available from the following website: <http://zju-capg.org/myo/data/index.html>). This dataset was developed for providing a standard benchmark database (DB) to explore new possibilities for studying next-generation muscle-computer interfaces (MCIs). The CapgMyo database comprises 3 sub-databases (referred to DB-a, DB-b and DB-c). However, DB-a has been used in our experiments to evaluate the performance of the proposed lightweight All-ConvNet for intra-session neuromuscular activity recognition because the maximum number of subjects (18) have participated in DB-a. In DB-a, 8 different isotonic and isometric hand gestures are performed by every subject and each hand gestures are recorded 10 times with a 1000 Hz sampling rate, which in total generates  $(8 \times 10 \times 1\,000) = 80\,000$  or 80k instantaneous sEMG images individually. Then, our All-ConvNet network is trained from scratch through *random initialization*. We performed training, validation and testing using only 80 000 images produced by 18 subjects individually through a leave one trial out cross-validation. We kept one trial out from each of the 8 different hand gestures i.e. 8 000 images for validation and testing. The remaining 9 trials for 8 different hand gestures i.e. 72k images are used for training. The cross-validation accuracy  $A$  is computed for each class  $i$ , as the number of totals correctly recognized hand gestures, divided by the total number of tests sEMG images

$$\text{Accuracy, } A = \frac{c}{N} = \frac{\sum C_i}{N} \quad (4)$$

where  $i = \{1, 2, \dots, G\}$  and  $G$  is the number of gesture classes.

In contrast, existing approaches (e.g. [12] and [14]) for instantaneous HD-sEMG image recognition used a total of  $(18 \times 40\,000) = 720\,000$  or 720k training images for pre-training, while 40 000 images from each of the subject are used separately for *fine-tuning*. Therefore, the existing approaches involve a total of  $(720\,000 + 40\,000) = 760\,000$  or 760k images only in the training process. Fig. 3 shows the total number of images are used during training for *pre-training + fine-tuning vs random initialization*.



Fig. 3. Total number HD-sEMG images seen during training, for pre-training + fine-tuning vs. random-initialization.

In our experiments, the proposed All-ConvNet described in Section III were trained on the CapgMyo DB-a datasets without any *pre-training* or data augmentations. In order to facilitate in GAP, we only enhance the input HD-sEMG image

size from  $16 \times 8$  to  $16 \times 16$  by horizontal mirroring. Unlike [12], this enhancement does not increase the learning parameters in the proposed All-ConvNet. The connection weights for All-ConvNet networks were *randomly initialized* using Xavier initialization scheme [18], [24] and were trained using Adam optimization algorithms [23] with a momentum decay and scaling decay are initialized to 0.9 and 0.999 respectively. In contrast to SGD, Adam is an adaptive learning rate algorithm; therefore, it requires less tuning of the learning rate hyperparameter. The learning rate of 0.001 is initialized to all our experiments. The smaller batches of 256 randomly chosen samples from the training dataset are fed to the network during consecutive learning iterations for all our experiments. We set a maximum of 100 epochs for training our All-ConvNet model. However, to avoid overfitting we have also applied early stopping in which the training process is interrupted if no improvements in validation loss are noticed for 5 consecutive epochs. The Batch normalization [19] is applied after the input and before each non-linearity. Dropout [20] was applied on all layers with probabilities 25%. The All-ConvNet model was trained on a workstation with an Intel Core, 3.60 (i7-4790) processor, 16GB RAM and an NVIDIA RTX 2080 Ti GPU. Each epoch was completed in approximately 5s. The test results for the All-ConvNet model are presented in Table II and compared with state-of-the-art methods. It is noteworthy that, the results in Table II are only compared with [12] because the same complex *fine-tuned and pre-trained* networks were subsequently employed in [14] and [30], though in [30] sparse channel sEMG were used.

As can be seen in Table II, the proposed All-ConvNet networks (on the order of only 460k learning parameters) consists of a stack of  $3 \times 3$  convolutional layers with occasional subsampling by stride of 2 and trained from *random initialization* performs comparably on CapgMyo DB-a dataset to the S-ConvNet [29] and fine-tuned pre-trained networks [12] even though the [12] use more complicated network architectures and training schemes which requires to learn over 5.63 millions parameters during fine-tuning only and also pre-trained with over 720k instantaneous HD-sEMG images. The average recognition accuracy of 8 hand gestures for 18 different subjects 85.81% obtained with the proposed All-ConvNet. The high average recognition accuracies 93.10%, 95.57% and 97.64% are achieved by a simple majority voting

TABLE II THE AVERAGE RECOGNITION ACCURACY (%) OF 8 HAND GESTURES WITH INSTANTANEOUS HD-SEM G IMAGES FOR 18 DIFFERENT SUBJECTS AND RECOGNITION APPROACHES. MAJORITY VOTING (ON 40 SEM G IMAGE) RESULTS ARE SHOWN IN PARENTHESSES

Model	Average Recognition Accuracy (%)	# Learning Parameters	Avg-run time (Sec.)
S-ConvNet [29]	87.95 (98.87)	$\approx 2\ 090k$	372.14
<b>All-ConvNet (proposed)</b>	<b>85.81 (97.64)</b>	<b><math>\approx 460k</math></b>	<b>348.54</b>
W.Geng <i>et al.</i> , [12]	89.3 (99.00)	$\approx 5\ 632k + \text{Pre-training}$	2091.39 (with no pre-training)

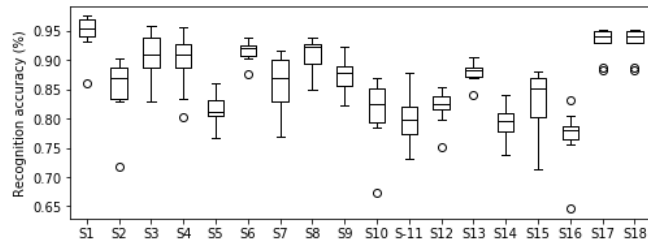


Fig. 4. The recognition accuracy of 8 hand gestures for 18 different subjects with our proposed All-ConvNet.

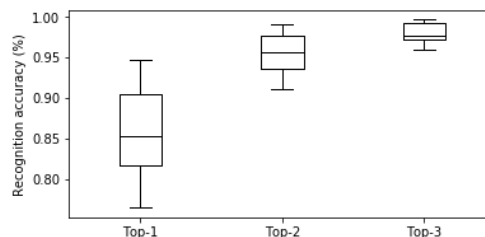


Fig.5 The proposed All-ConvNet - Top-K recognition accuracy (K = 1, 2, 3) with 3, 20 and 40 instantaneous images respectively. The average run time of training, validation and predictions for an intra-subject test is also included in Table II. The proposed All-ConvNet outperformed the existing methods.

The recognition accuracy of 8 hand gestures of all 18 subjects in CapgMyo DB-a which obtained through leave one trial out cross validation for 10 different trials using All-ConvNet and their statistical significance are presented in Fig. 4. It is observed that the average recognition accuracy  $>92.21\%$  and  $>87.36\%$  have been achieved at least for 6 and 5 different subjects respectively.

Perhaps even more interesting, the proposed All-ConvNet achieved the state of the art recognition accuracy when the instantaneous HD-sEMG images are recognized by Top-2 or Top-3 performance metrics i.e. when the target gestures (neuromuscular activities) are matched to any of the 2 or 3 highest probabilities provided by the SoftMax layer of the All-ConvNet. Fig. 5 presents the Top-1, Top-2 and Top-3 accuracies respectively. The obtained average Top-2 and Top-3 recognition accuracies are 95.60% and 98.07% respectively. These outstanding results confirm that the proposed lightweight All-ConvNet trained from *random initialization* is highly effective for learning all the invariances for low-resolution instantaneous HD-sEMG image recognition and hence seem to be enough to address the aforementioned problem of employing high-end resource bounded fine-tuned pre-trained networks for low-resolution instantaneous HD-sEMG image recognition.

The existing neuromuscular activity recognition methods [12], [14] require a huge memory space to store the massive parameters. Therefore, the models are usually unsuitable for low-end hand-held devices and embedded electronics. Thanks to the proposed parameter-efficient All-ConvNet, our model is much smaller than the most competitive methods for instantaneous HD-sEMG image recognition.

## V. CONCLUSION

We present a lightweight All-ConvNet network, a simple yet efficient framework for learning instantaneous HD-sEMG images from scratch for neuromuscular activity recognition. Without using any pre-trained models, our proposed All-ConvNet demonstrates very competitive or state of the art performance, while using  $\approx 12 \times$  smaller dataset and reducing learning parameters from  $\approx 5.63 M$  to only  $\approx 460 k$  than its fine-tuned pre-trained counterparts for neuromuscular activity recognition based on instantaneous HD-sEMG images. The proposed All-ConvNet has great potential for learning and recognizing neuromuscular activities on resource-bounded devices. Our future work will consider improving inter-session neuromuscular activity recognition performances, as well as learning All-ConvNet models to support resource-bounded devices.

## ACKNOWLEDGMENT

This work was supported in part by the Regroupement stratégique en microsystèmes du Québec (ReSMiQ) and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

## REFERENCES

- [1] D. Farina *et al.*, "The extraction of neural information from the surface EMG for the control of upper-limb prostheses: merging avenues and challenges," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 2, pp. 797–809, 2014.
- [2] G. Jang, J. Kim, J. S. Lee, and Y. Choi, "EMG-based continuous control scheme with simple classifier for electric-powered wheelchair," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 6, pp. 3695–3705, 2016.
- [3] R. Jimenez-Fabian, and O. Verlinden, "Review of control algorithms for robotic ankle systems in lower-limb orthoses, prostheses, and exoskeletons," *Medical Engineering & Physics*, vol. 34, no. 4, pp. 397–408, 2012.
- [4] Y. Hu, J. N. Mak, & K. Luk, "Application of surface EMG topography in low back pain rehabilitation assessment," *International IEEE/EMBS Conference on Neural Engineering*, pp. 557–560, 2007.
- [5] E. Costanza, S. A. Inverso, R. Allen, and P. Maes, "Intimate interfaces in action: assessing the usability and subtlety of EMG-based motionless gestures," *Conference on Human Factors in Computing Systems*, ACM, pp. 819–828, 2007.
- [6] T. S. Saponas, D. S. Tan, D. Morris, and R. Balakrishnan, "Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces," *Conference on Human Factors in Computing Systems*, ACM, pp. 515–524, 2008.
- [7] T. S. Saponas, D. S. Tan, D. Morris, D. J. Turner and J. A. Landay, "Making muscle-computer interfaces more practical," *Conference on Human Factors in Computing Systems*, pp. 851–854, ACM, 2010.
- [8] M. Atzori *et al.*, "Electromyography data for non-invasive naturally-controlled robotic hand prostheses," *Scientific Data* 1, 2014.
- [9] N. Patricia, T. Tommasi, & B. Caputo, "Multi-source adaptive learning for fast control of prosthetics hand," *International Conference on Pattern Recognition*, pp. 2769–2774, 2014.
- [10] C. Amma, T. Krings, J. Ber, J. and T. Schultz, "Advancing muscle-computer interfaces with high-density electromyography," *Conference on Human Factors in Computing Systems*, pp. 929–938, ACM, 2015.
- [11] A. Stango, F. Negro and D. Farina, "Spatial correlation of high density EMG signals provides features robust to electrode number and shift in pattern recognition for myocontrol," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* vol. 23, no.2, pp. 189–198, 2015.
- [12] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu and J. Li, "Gesture recognition by instantaneous surface EMG images," *Scientific Reports.*, vol. 15, no.6, 36571, Nov 2016.
- [13] M. R. Islam, D. Massicotte, F. Nougrou and W. Zhu, "HOG and pairwise SVMs for neuromuscular activity recognition using instantaneous HD-sEMG images," *In IEEE NEWCAS*, pp. 335-339, 2018.
- [14] Y. Du., W. Jin, W. Wei, Y. Hu and W Geng, "Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation," *Sensors*, vol. 17, no. 3, 458, 2017.
- [15] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: closing the gap to human-level performance in face verification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701-1708, 23-28 June 2014.
- [16] L. Pang, Y. Lan, J. Xu, J. Guo, and X. Cheng., "Locally smoothed neural networks," *arXiv preprint arXiv:1711.08132*, 2017.
- [17] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller., "Striving for simplicity: the all convolutional net," *CoRR*, abs/1412.6806, 2014.
- [18] X. Glorot and Y. Bengio., "Understanding the difficulty of training deep feedforward neural networks," *In AISTATS*, 2010.
- [19] S. Ioffe and C. Szegedy., "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *In ICML*, 2015.
- [20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp.1929–1958, 2014.
- [21] M. Lin, Q. Chen, and S. Yan, "Network in network," *In ICLR: Conference Track*, 10 pages, 2014.
- [22] K. Janocha, and W. M. Czarnecki. "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.
- [23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [24] K. He, X. Zhang, S. Ren, and J. Sun., "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," *In ICCV*, 2015.
- [25] D.-A. Clevert, T. Unterthiner, and S. Hochreiter., "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.
- [26] F. Nougrou, A. Campeau-Lecours, M. R. Islam, D. Massicotte and B. Gosselin, "Muscle activity distribution features extracted from HD sEMG to perform forearm pattern recognition," *In IEEE LSC*, pp. 275-278, 2018.
- [27] F. Nougrou, A. Campeau-Lecours, D. Massicotte, M. Boukadoum, C. Gosselin, and B. Gosselin. "Pattern recognition based on HD-sEMG spatial features extraction for an efficient proportional control of a robotic arm." *Biomedical Signal Processing and Control* 53 (2019): 101550.
- [28] U. Côté-Allard *et al.*, "Deep learning for electromyographic hand gesture signal classification using transfer learning,"

in *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 760-771, April 2019.

- [29] M. R. Islam, D. Massicotte, F. Nougrou, P. Massicotte and W. Zhu., "S-ConvNet: A shallow convolutional neural network architecture for neuromuscular activity recognition using instantaneous high-density surface EMG images," *arXiv preprint arXiv:1906.03381*, 2019.
- [30] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli and W. Geng, "Surface-electromyography-based gesture recognition by multi-view deep learning," in *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2964-2973, Oct. 2019.



M. R. Islam, D. Massicotte, P. Massicotte and W-P Zhu, "Surface EMG-Based Inter-Session/Inter-Subject Gesture Recognition by Leveraging Lightweight All-ConvNet and Transfer Learning," *IEEE Transactions on Instrumentation and Measurements*, 2024 (Accepted).

# Surface EMG-Based Inter-Session/Inter-Subject Gesture Recognition by Leveraging Lightweight All-ConvNet and Transfer Learning

Md. Rabiul Islam, *Student Member, IEEE*, Daniel Massicotte, *Senior Member, IEEE*,  
Philippe Y. Massicotte, and Wei-Ping Zhu, *Senior Member, IEEE*

**Abstract**— Gesture recognition using low-resolution instantaneous high-density surface electromyography (HD-sEMG) images opens up new avenues for the development of more fluid and natural muscle-computer interfaces. However, the data variability between inter-session and inter-subject scenarios presents a great challenge. The existing approaches employed very large and complex deep ConvNet or 2SRNN-based domain adaptation methods to approximate the distribution shift caused by these inter-session and inter-subject data variability. Hence, these methods also require learning over millions of training parameters and a large pre-trained and target domain dataset in both the pre-training and adaptation stages. As a result, it makes high-end resource-bounded and computationally very expensive for deployment in real-time applications. To overcome this problem, we propose a lightweight All-ConvNet+TL model that leverages lightweight All-ConvNet and transfer learning (TL) for the enhancement of *inter-session* and *inter-subject* gesture recognition performance. The *All-ConvNet+TL* model consists solely of convolutional layers, a simple yet efficient framework for learning invariant and discriminative representations to address the distribution shifts caused by inter-session and inter-subject data variability. Experiments on four datasets demonstrate that our proposed methods outperform the most complex existing approaches by a large margin and achieve state-of-the-art results on inter-session and inter-subject scenarios and perform on par or competitively on intra-session gesture recognition. These performance gaps increase even more when a tiny amount (e.g., a single trial) of data is available on the target domain for adaptation. These outstanding experimental results provide evidence that the current state-of-the-art models may be overparameterized for sEMG-based inter-session and inter-subject gesture recognition tasks.

**Index Terms**— Transfer learning, domain adaptation, convolutional neural network, recurrent neural network, feature extraction, muscle-computer interface, surface electromyography, EMG, gesture recognition

## I. INTRODUCTION

GESTURE recognition based on surface electromyography (sEMG) signals has been a core technology for developing next-generation muscle-computer interfaces (MCIs). The major application domains of sEMG-based MCIs are non-intrusive control of active prosthesis [1], wheelchairs [2], exoskeletons

[3] or neurorehabilitation [4], neuromuscular diagnosis [5] and providing interaction methods for video games [6], [7]. The existing approaches for gesture recognition using sparse multi-channel sEMG sensors and classical machine learning methods – such as linear discriminant analysis (LDA) [8], support vector machines (SVM) [9], hidden Markov model (HMM) [10] – on windowed descriptive and discriminative time-domain, frequency-domain and/or time-frequency-domain sEMG feature space [11], [12-16]. However, these sparse multi-channel sEMG-based methods are not suitable for real-world applications due to their lack of robustness to electrode shift and positioning [17], [18]. In addition, malfunction to any of these sparse-channel electrodes leads to retraining the entire MCI system. Deep learning-based methods have recently been exploited for gesture recognition using sparse multi-channel sEMG [19-20], [31-32], [61], [68] but their performance is still far from optimum [64].

To address this problem, designing and developing more flexible, convenient, and comfortable high-density sEMG (HD-sEMG) based myoelectric sensors and efficient pattern recognition algorithms have been major research directions in recent years [17-18], [21-30], [36]. However, the existing HD-sEMG-based gesture recognition methods [17-18], [28], [30] still rely on the windowed sEMG (e.g., range between 100 ms and 300 ms [33], [34]), which demands finding an optimal window length. The determination of an optimal window length represents a strong trade-off between classification accuracy and controller delay, both of which increase with an increase in window size.

To further address this problem, distinctive patterns within instantaneous sEMG images were first discovered by Geng et al. [21] and M.R. Islam et al. [22] to develop more fluid and natural muscle-computer interfaces (MCIs). The instantaneous values of HD-sEMG signals at each sampling instant were arranged in a 2D grid in accordance with the electrode positioning. Subsequently, this 2D grid was transformed into a grayscale sEMG image. Therefore, an instantaneous sEMG image represents a relative global measure of the physiological

Md. R. Islam, D. Massicotte, and P.Y. Massicotte are with the Laboratory of Signal and System Integration (LSSI), Department of Electrical and Computer Engineering, Université du Québec à Trois-Rivières, Trois-Rivières, QC, G9A 5H7, Canada. (e-mail: md.rabiul.islam@uqtr.ca; daniel.massicotte@uqtr.ca; philippe.massicotte2@uqtr.ca).

W.-P. Zhu is with the Department of Electrical and Computer Engineering,

Concordia University, Montreal, H3G 1M8, Canada. (e-mail: weiping@ece.concordia.ca).

This work has been funded by the Natural Sciences and Engineering Research Council of Canada grant, CMC Microsystems, and the Research Chair on Signal and Intelligent high-performance systems.

processes underlying neuromuscular activities at a given time. Consequently, gesture recognition is performed solely with the sEMG images spatially composed from HD-sEMG signals recorded at a specific instant.

Motivated by these prior works, further studies have been conducted on this promising new research direction over the years [23-27], [29], [36]. However, the state-of-the-art methods [21], [23], [24] for sEMG-based gesture recognition either employed very complex deep and wide CNN or an ensemble of these complex networks for improved gesture recognition performance. Despite the significant performance boost achieved by these state-of-the-art models [21], [23], [24], the heavy computational and intensive memory cost hinders deploying them on resource-constrained embedded and mobile devices for real-time applications.

In addition, the sEMG-based gesture recognition problem becomes more challenging in the operational conditions or an *inter-session* scenario, where the trained model is used to recognize muscular activities in a new recording session because sEMG signals are highly subject-specific. The distributions of the sEMG signals vary considerably even between recording sessions of the same subject within the same experimental setup. The acquired sEMG signals in a new recording session (target domain or task) differ from those obtained during the training session (source domain or task) because of electrode shifts, changes in arm posture, and slow time-dependent changes such as fatigue and electrode-skin contact impedance [1][26]. Inter-session is often referred to as inter-subject when the training and test data are acquired from different subjects. Moreover, it is always challenging to force the users to maintain a certain level of muscular contraction force in real-time applications. Therefore, the developed methods must also cope with the distribution shift occurred by this voluntary muscular contraction force level.

To attenuate these distribution shifts between different sEMG recording sessions, the *pre-trained models* have been predominantly adopted by the existing approaches [26], [31], [32], and [57] to reduce the distribution shift by *fine-tuning* the sEMG data recorded in the different session (target domain or task). *Fine-tuning* updates the parameters of the *pre-trained models* to train to newly recorded sEMG data. Generally, the output layer of the pre-trained models is extended with randomly initialized weights. A small learning rate is used to *fine-tune* all the parameters from their original values to minimize the loss on the newly recorded sEMG data. Using appropriate hyper-parameters for training, the resulting *fine-tuned model* often outperforms learning from a randomly initialized network [40].

Generally, this *pre-training* and *fine-tuning* process can be considered a special case of *domain adaptation* when the *source task* and the *target task* are the same or *transfer learning* when the tasks are different. However, for sEMG-based gesture recognition scenarios, we reframed this problem as *transfer learning* when the sEMG data for training and inference are recorded at a different session. Fig. 1 illustrates the conceptual diagram of our proposed transfer-learning methods for

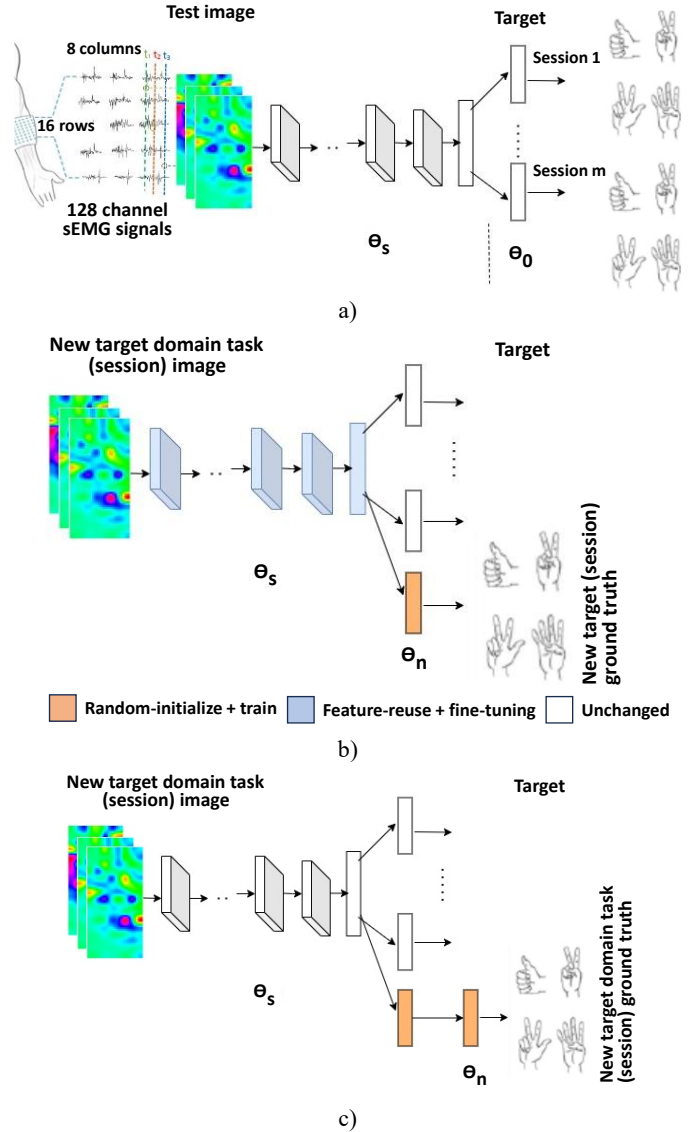


Fig. 1. A general conceptual diagram of the transfer learning method (a) Pre-trained model (b) Fine-tuned model and (c) Feature extraction process. sEMG images and labels used for adaptation are shown.

sEMG-based gesture recognition.

Transfer learning is typically performed by taking a standard architecture along with its pre-trained weights and then *fine-tuning* the target task. However, the state-of-the-art methods [21], [23], [26], and [61] for sEMG-based gesture recognition employed very large and deep pre-trained models, therefore, containing millions of parameters which are designed to be trained with large-scale labeled sEMG datasets. The requirement of high-end computing resources and large-scale *pre-trained* datasets are also bounded by large and deep network structures [25]. As far as we are aware, there has been no research for sEMG-based gesture recognition studying the effects of transfer learning on the smaller, simpler, and lightweight CNN. This line of investigation is especially crucial in the sEMG-based gesture recognition because the pre-trained model is often deployed in real-time MCI applications such as assistive technology and physical rehabilitation where fine-tuning in the target domain must be conducted in the data-

starved condition because of the difficulty of acquiring data from the amputees, elderly peoples, and patients, etc. Also, the large computationally expensive models might significantly impede mobile and on-device applications, where power consumption, data memory, and computational speed are constraints. To investigate the effects of transfer learning for sEMG-based gesture recognition, our research is motivated by the following research questions- *does feature reuse takes place during fine-tuning or transfer learning? And if yes, where exactly is it in the network?*

Investigating feature reuse, we find out that some of the differences from transfer learning are due to the over-parametrization of the state-of-the-art, more complex pre-trained models rather than sophisticated feature reuse. Additionally, we discovered that a simple, lightweight model can outperform the more complex and computationally demanding state-of-the-art network architectures. We isolate where useful feature reuse occurs and outline the implications for more efficient lightweight model exploration.

In this paper, we perform a fine-grained study on fine-tuning and transfer learning for sEMG-based gesture recognition. Our main contributions are:

- (1) We introduce All-ConvNet+TL model, which leverages the lightweight All-ConvNet and transfer learning to address the distribution shift in inter-session and inter-subject sEMG-based gesture recognition and evaluate it against the more complex state-of-the-art network architectures. Our proposed method leveraging lightweight All-ConvNet and transfer learning outperforms the state-of-the-art methods by a large margin, both when the data from a single trial or multiple trials are available for *fine-tuning/adaptation*. The outstanding *inter-session* and *inter-subject* gesture recognition performance achieved by the proposed lightweight models raises the question of whether the current state-of-the-art models are overparameterized for the sEMG-based gesture recognition problem.
- (2) Using further analysis and weight transfusion experiments, where we partially reuse pre-trained weights, we identify locations where meaningful feature reuse occurs and explore hybrid approaches to transfer learning. These approaches involve using a subset of pre-trained weights and redesigning other parts of the network to make them more lightweight.
- (3) We conducted more extensive experiments. A performance evaluation on CapgMyo and its four (4) publicly available HD-sEMG sub-datasets was performed on three different sEMG-based gesture recognition tasks: *intra-session*, *inter-session*, and *inter-subject* scenarios. The results showed that our lightweight models outperformed the more complex state-of-the-art models on various tasks and datasets.

The rest of the paper is structured as follows: Section II reviews current state-of-the-art methods for sEMG-based gesture recognition, Section III presents the proposed transfer learning framework, while Section IV presents the lightweight All-ConvNet model architecture and its design principles. Section V introduces the proposed transfer learning design

methodology by leveraging lightweight All-ConvNet (All-ConvNet+TL). Section VI describes the experimental framework, and Section VII demonstrates the state-of-the-art results for inter-session and inter-subject gesture recognition and very competitive results for intra-session gesture recognition, obtained from experiments conducted on CapgMyo and its four (4) sub-datasets. Section VIII highlights the state-of-the-art performance achieved by the proposed All-ConvNet+TL and discusses some important findings. Finally, Section IX provides some conclusive remarks.

## II. RELATED WORK

In this section, we present an overview of current state-of-the-art methods for sEMG-based gesture recognition. Many efforts have been devoted to proposing novel deep learning methods to enhance the accuracy of sEMG-based gesture recognition. Geng et al. [21] employed a deep convolutional neural network (CNN or ConvNet) to recognize hand gestures from the sEMG images and showed high recognition accuracy on publicly available benchmark HD-sEMG datasets [15], [17], [26]. M. R. Islam et al. [22] proposed to use Histogram of Oriented Gradients (HoG) as discriminative features and an SVM-based feature classification algorithm for high-density EMG images, achieving accurate classification of 8 gestures [11]. Motivated by [21] and [22], further studies have been conducted in recent years [23-27], [29], [36]. Wei et al. [23] proposed a two-stage convolutional neural network (CNN) with a multi-stream decomposition stage and a fusion stage to learn the correlation between certain muscles and specific gestures. The sEMG image is decomposed into different equally sized image patches based on the layout of the electrode arrays on muscles (e.g., each of eight  $8 \times 2$  electrode arrays in the CapgMyo database [26] individually produces  $8 \times 2$  equal-sized sEMG image patches). Then, each of these sEMG image patches is independently and in parallel passed through the convolution layers of a single-stream CNN [21], thereby forming a multi-stream CNN. The learned features from all the single-stream CNNs that form a multi-stream CNN are aggregated and fed to a fusion network for gesture recognition. The reported results showed that multi-stream CNN outperformed single-stream CNN by a small margin. Hu et al. [24] proposed a combined CNN-RNN module to capture both spatial and temporal information of sEMG signals for gesture recognition. The recorded sEMG signals were decomposed into small subsegments using a sliding and overlapping windowing strategy. Each of these sEMG subsegments was converted into an sEMG image and simultaneously passed through a multi-stream CNN built upon [21] for feature extraction. Given the input sequence of the extracted features corresponding to each of the sEMG subsegments, a long short-term memory (LSTM) network was learned individually for gesture recognition. Then, the features learned by each of these LSTMs corresponding to each of these sEMG subsegments were concatenated before being fed to a fully connected and SoftMax layer for gesture recognition. Experimental results indicate that a combined CNN-RNN module outperforms the stand-alone CNN and

RNN frameworks, respectively. Encouraged by [38], Chen et al. proposed to use of 3D convolution in the convolutional layers of CNNs for spatial and temporal representation of sEMG images [36]. The 3D convolution is attained by convolving a 3D kernel to the cube formed by stacking multiple adjacent sEMG image frames. The feature maps in the convolution layers of a 3D CNN are connected to multiple adjacent sEMG image frames in the previous layer. Hence, the spatiotemporal information is captured. However, multiple 3D convolutions with distinct kernels are required to apply at the same location of the input to learn representative features, which makes 3D CNN computationally expensive. For example, the exploited 3D CNN in [36] requires learning over >30M (million) parameters when the length of the input cube is set to 10 (i.e., the cube is formed by stacking 10 consecutive sEMG image frames).

However, current state-of-the-art methods [21], [23], [24] employed complex deep and wide CNNs or network ensembles for enhanced gesture recognition performance. For example, Geng et al. [21] exploited a DeepFace [35] like very large and deep CNN (dubbed as GengNet), which requires learning >5.63M (million) training parameters only during *fine-tuning* and *pre-trained* on a very large-scale labeled sEMG training datasets. The complexity of this model grows linearly as the input size is increased due to the use of an unshared weight strategy [27]. Wei et al. [23] used an ensemble of eight (8) single-stream GengNet at the decomposition stage only. Hu et al. [24], used a two-stage ensemble network in which an ensemble of multiple single-stream GengNet was used for spatial feature learning, resulting in multiple sequences of 1-D feature representation. Then, these 1-D feature sequences were passed to an ensemble of LSTM networks before a SoftMax layer recognized the targeted gesture. Hence, deploying these state-of-the-art models [21], [23], and [24] on embedded and mobile devices for real-time applications becomes cumbersome, despite achieving significant performance gains. Therefore, the demand for designing low-cost, lightweight networks is highly increasing for low-end resource-limited embedded and mobile devices.

To overcome these problems, more recently, low-latency and parameter-efficient S-ConvNet [25] and All-ConvNet [27] have been introduced, targeting sEMG-based gesture recognition on low-end devices. S-ConvNet [25] was designed to learn sEMG image representation *from scratch through random initialization*. S-ConvNet consists of a network with convolution layers with the shared kernel, a fully connected layer with a small number of neurons, and an occasional dimensionality reduction performed by stridden CNN, demonstrating very competitive gesture recognition accuracy while needing to be learnt  $\approx 1/4$ th learning parameters using a  $\approx 12 \times$  smaller dataset compared to the more complex and high-end resource-bounded state-of-the-art [21]. A similar CNN architecture to that of S-ConvNet is used by Tam et al. [29] for a fully embedded adaptive real-time sEMG-based gesture recognition. Striving to find a simpler and more efficient lightweight network, in our recent work [27], a new

architecture called All-ConvNet was introduced that consists solely of convolutional layers and is designed to be more efficient and less computationally intensive than the existing state-of-the-art models for sEMG-based gesture recognition. Comparing the performance of All-ConvNet to other state-of-the-art models shows that it achieves competitive or state-of-the-art performance on a current benchmark HD-sEMG dataset [26], while being significantly lighter, more efficient, and faster to train and evaluate. *All-ConvNet was designed based on the finding of fact that if the sEMG image area covered by units in the topmost convolutional layer covers a portion of the image large enough to recognize its content (i.e., gesture class we want to recognize)*. This leads to predictions of sEMG image classes at different positions which can then simply be averaged over the whole image. Hence, the All-ConvNet becomes robust to translations and geometric distortions, which can be very effective in addressing the electrode shift and positioning problem in sEMG-based gesture recognition.

Moreover, *pre-trained* models have been employed by [26], [31], [32], and [57] to mitigate distribution shifts by *fine-tuning* on the target domain or task for sEMG-based gesture recognition in inter-session and inter-subject scenarios. Currently, Du et al. [26] and Ketyko et al. [57] present state-of-the-art solutions for sEMG-based gesture recognition in inter-session and inter-subject scenarios. Du et al. [26] propose a multi-source extension to the classical adaptive batch normalization (AdaBN) technique [37], combined with their most complex deep and large CNN architecture [21]. They employ AdaBN with the hypothesis that the layer weights contain discriminative knowledge related to different hand gestures, while the statistics of the BatchNorm layer [55] represent discriminative knowledge from different recording sessions in inter-session or inter-subject scenarios [37]. The parameters of the pre-trained model's AdaBN [21] are updated using an unsupervised approach for adaptation in the target domain. However, a drawback of this solution arises when dealing with multiple sources (i.e., multiple subjects), as specific constraints and considerations must be imposed for each source during the pre-training phase of the model [57]. Ketyko et al. [57] proposed a 2-Stage recurrent neural networks (2SRNN), where a deep stacked RNN sequence classifier was used for pre-training on the source dataset. Then, the weights of the pre-trained deep-stacked RNN classifier were frozen. At the same time, a fully connected layer without a non-linear activation function was trained in a supervised manner on the target dataset for domain adaptation. More explicitly, the deep-stacked RNN classifier was used as a feature extractor by freezing its weight in the domain adaptation stage. However, ConvNet is computationally more efficient and powerful in extracting discriminative features than RNN, even for classification tasks involving long sequences [58], [59]. Unlike these works, the proposed All-ConvNet+TL model capitalizes the inherent invariant properties of translations and geometric distortions in All-ConvNet and investigates the feasibility of applying transfer learning (TL) on the smaller, simpler, and lightweight All-ConvNet to address the distribution shift and

learn invariant, discriminative representations for efficient sEMG-based gesture recognition in inter-session and inter-subject scenarios.

### III. THE PROPOSED TRANSFER LEARNING FRAMEWORK

The proposed transfer learning framework for sEMG-based gesture recognition using instantaneous HD-sEMG images includes the following three major computational components: (i) a *lightweight model development* (ii) *pre-training*, and (iii) *fine-tuning*. A schematic diagram of the proposed transfer learning framework for sEMG-based gesture recognition is shown in Fig. 1. Firstly, we devised a lightweight All-ConvNet model. Secondly, the proposed lightweight All-ConvNet was pre-trained (e.g., Fig. 1a) using a large amount of gesture data acquired by HD-sEMG in a single session or over multiple sessions, which may also involve multiple gestures, trials, and subjects, respectively. Then, the pre-trained model was saved and deployed for subject-specific/personalized classifier development, as sEMG-based wearable devices are usually worn by a single user while executing a target task. Typically, input-side layers that play the role of feature extraction are copied from a pre-trained network and kept frozen or fine-tuned (e.g., Fig. 1b and 1c), in contrast, a top classifier for the target task is randomly initialized and then trained at a slow learning rate. *Fine-tuning* often outperforms training from scratch because the pre-trained model already has a great deal of muscular activity information. Potentially, the pre-trained network could be duplicated and fine-tuned for each new target task [40].

### IV. MODEL DESCRIPTION – THE ALL-CONVOLUTIONAL NEURAL NETWORK (ALL-CONVNET)

The current state-of-the-art methods [21], [23], [26], and [61] for sEMG-based gesture recognition use a large, deep ConvNet architecture similar to the one used in DeepFace [35]. This architecture is designed to be pre-trained on a large-scale labeled HD-sEMG training dataset and requires learning  $>5.63$  million (M) parameters only during fine-tuning. As a result, this large-scale pre-trained model becomes a high-end resource-bounded and computationally very expensive to be practical for real-world MCI applications. Moreover, in their pre-trained ConvNet includes two locally connected (LCN) and three fully connected layers among the other convolutions and a  $G$ -way fully connected layer. However, the LCN layers used an unshared weight scheme [45] that makes their pre-trained ConvNet even computationally more demanding and very difficult to scale on the target domain task. For example, the learning parameters of [21] increase from  $\approx 5.63$ M to  $\approx 11$ M with a small enhancement of input HD-sEMG image size from  $16 \times 8$  to  $16 \times 16$  due to the use of this unshared weight scheme [27]. Hence, a very large-scale labeled training dataset is required for learning these growing numbers of training parameters [35]. However, the LCN can be beneficial in the application domains where the feature’s precise location is dependent on the class labels.

Considering the above-mentioned fact, we investigated the following research questions in [27] : (i) Do we expect the devised networks model to produce a location/translation

invariant feature representation? and (ii) Do we need a location-dependent feature representation? Following our findings and building on other recent works that aim to find a simple network architecture, we proposed a lightweight All-ConvNet. This new architecture consists solely of convolutional layers. This simple yet effective framework could learn neuromuscular activity from scratch and yield competitive or even state-of-the-art performance using a  $\approx 12 \times$  smaller dataset while reducing the learning parameters from  $\approx 5.63$ M to only  $\approx 460$ k than the more complex state-of-the-art for sEMG-based gesture recognition. The All-ConvNet architectural design was adopted based on the following principles and observations:

- (i) We hypothesized that different hand gestures produce distinct spatial intensity distributions that remain consistent across multiple trials of the same gesture and distinguishable among different gestures. However, we observed that the spatial intensity distributions for the same gesture are not locally invariant, and the precise feature’s location are independent of the class labels. Fig. 2 demonstrates a sequence of HD-sEMG images derived from the same class, along with a correlation heatmap of HD-sEMG distributions (images) sampled equidistantly in time (e.g., each 20 ms) which demonstrates that the distributions are independent of the class labels. CNN alone has a remarkable capability to exploit locally translational invariance features by utilizing local connectivity and weight-sharing strategies [45]. On the other hand, the LCN layer fails to model the relations of parameters in different locations. Hence, the LCN layers are ablated in designing

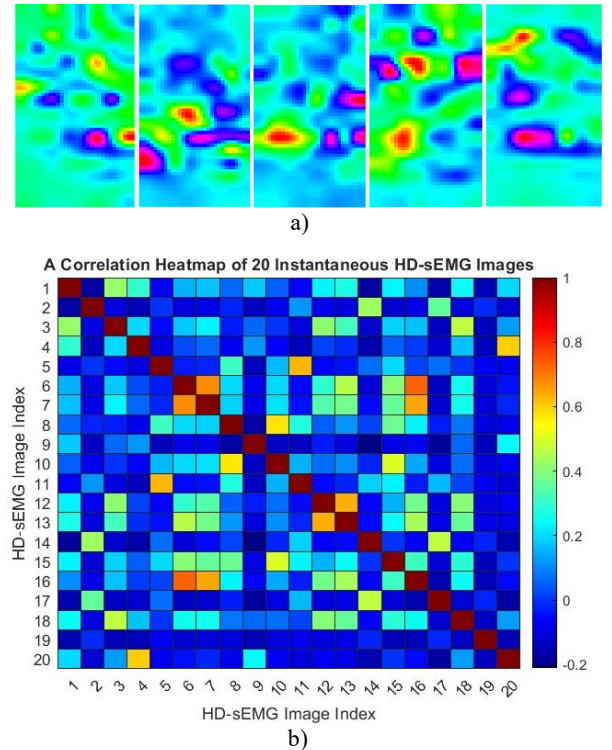


Fig. 2 HD-sEMGs derived from the same muscular activity class (a) and correlation heatmap of HD-sEMG distributions (b) which demonstrates that the distributions are independent to the class labels.

our All-ConvNet models as the location of the features is not dependent on the class labels.

- (ii) Inspired by previous work [46], we leverage the fact that if the part of the instantaneous HD-sEMG image is covered by the units in the topmost convolution layers could be large enough to recognize its content (i.e., the gesture class, we want to recognize). Consequently, the fully connected layers can also be replaced by simple 1-by-1 convolutions. This allows us to predict HD-sEMG image classes at different positions, and we can then average these predictions across the entire image. Hence, the proposed All-ConvNet can be very effective in addressing the electrode shift and positioning problem for sEMG-based gesture recognition, where the entire sEMG data stream for a particular gesture may not necessarily be required for recognition. Lin et al. [47], initially introduced this approach, which acts as an additional regularization technique due to the significantly fewer parameters of a 1-by-1 convolution in comparison to a fully connected and LCN layers. Overall, our architecture is thus reduced to consist only of convolutional layers with ELU nonlinearities [48], [63] and a global average pooling (GAP) + SoftMax layer to produce predictions over the entire instantaneous HD-sEMG image. A conceptual diagram of our proposed pre-trained All-ConvNet is shown in Fig. 1(a). Table I describes our proposed All-ConvNet architecture. The feature maps learned by the proposed All-ConvNet are presented in Fig. 3.

We train our proposed All-ConvNet for a multi-class sEMG-based gesture recognition task, which involves recognizing a specific muscular activity class using an instantaneous HD-sEMG image. As described in Table I, in the proposed All-ConvNet network, we consider using 1-by-1 convolution at the top to produce 8 or 12 outputs (depending on the number of distinct movements performed). These outputs were then averaged across all positions and fed into a  $G$ -way SoftMax layer (where  $G$  is the number of distinct hand gesture classes) which produces a distribution over the class labels. In order to estimate the class probabilities, we use the SoftMax function  $\sigma(\cdot)$  with  $\hat{y}^{(j)}$  representing the  $j$ th element of the  $G$  dimensional output vector of the layer preceding the SoftMax layer, defined as below:

$$\sigma(\hat{y}^{(j)}) = \frac{\exp(\hat{y}^{(j)})}{\sum_G \exp(\hat{y}^{(G)})} \quad (1)$$

The objective of this training is to maximize the probability of the correct gesture class. This is accomplished by minimizing the cross-entropy loss [49] for each training sample. When  $y$  represents the true label for a given input, the loss is computed as:

$$L = -\sum_j y^{(j)} \ln(\sigma(\hat{y}^{(j)})) \quad (2)$$

The loss is minimized over the parameters by computing the gradient of  $L$  with respect to the parameters. These parameters are then updated using the state-of-the-art Adam (adaptive moment estimation) gradient descent-based optimization algorithm [50]. This algorithm provides fast and reliable learning convergence, unlike the stochastic gradient descent (SGD) optimization algorithm used in state-of-the-art pre-

TABLE I THE ALL-CONVNET NETWORK MODEL FOR NEUROMUSCULAR ACTIVITY RECOGNITION.

<b>All-ConvNet</b>	
Input $16 \times 16$ Gray-level Image	
$3 \times 3$ Conv. 64 ELU	
$3 \times 3$ Conv. 64 ELU	
$3 \times 3$ Conv. 64 ELU with stride $r=2$	
$3 \times 3$ Conv. 128 ELU	
$3 \times 3$ Conv. 128 ELU	
$3 \times 3$ Conv. 128 ELU with stride $r=2$	
$1 \times 1$ Conv. 128 ELU	
$1 \times 1$ Conv. 8 ELU	
global averaging over $4 \times 4$ spatial dimensions	
G-way SoftMax	

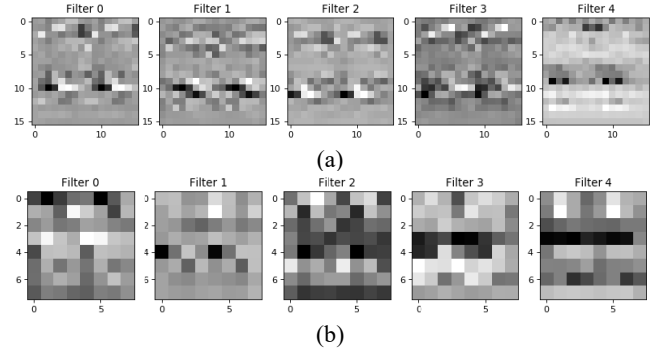


Fig. 3. A schematic illustration of feature maps obtained by All-ConvNet before and after dimensionality reduction. (a) Feature maps and b) Feature maps after dimensionality reduction.

trained networks for gesture recognition using instantaneous HD-sEMG image recognition.

Once the network has been trained, an instantaneous HD-sEMG image is recognized as in the gesture class  $\mathcal{C}$  by simply propagating the input image forward and computing:

$$\mathcal{C} = \operatorname{argmax}_j(\hat{y}^{(j)}) \quad (3)$$

## V. TRANSFER LEARNING BY LEVERAGING LIGHTWEIGHT ALL-CONVNET (ALL-CONVNET+TL)

In this section, we introduce some notations and definitions used in our transfer learning framework as in [51]. We denote the *source domain* data as  $D_s = \{(x_{s_1}, y_{s_1}), \dots, (x_{s_{n_s}}, y_{s_{n_s}})\}$ , where  $x_{s_i} \in X_s$  is the data instance and  $y_{s_i} \in Y_s$  is the corresponding class label. In our sEMG-based gesture recognition example,  $D_s$  can be a set of sEMG data of different gestures and their corresponding gesture class labels acquired by a single or multiple participants in a designated session. An objective function  $f_s(\cdot)$  can be learned using  $D_s$  for the source task such that,  $\mathcal{T}_s = \{Y_s, f_s(\sum_i w_{s_i} X_s + b)\}$ . Similarly, we denote the *target domain* data as  $D_T = \{(x_{T_1}, y_{T_1}), \dots, (x_{T_{n_T}}, y_{T_{n_T}})\}$  and  $\mathcal{T}_T = \{Y_T, f_T(\sum_i w_{T_i} X_T + b)\}$ , where,  $x_{T_i} \in X_T$  and  $y_{T_i} \in Y_T$  are the sEMG data of different gestures and their corresponding class labels respectively acquired by a distinct subject/participant at a different session than  $D_s$ . In most cases, the *target domain data* for a distinct participant acquired at another session is much lower quantities than that of a *source domain data*, i.e.  $0 \leq n_T \ll n_s$ .

Now we define our proposed transfer learning problem as follows— Given a source domain  $D_s$  and a learning task  $\mathcal{T}_s$  as well as a target domain  $D_T$  and learning task  $\mathcal{T}_T$ , the transfer learning aims to help improve the learning of the target predictive function  $f_T(\cdot)$  in  $D_T$  using the knowledge in  $D_s$  and  $\mathcal{T}_s$ , where,  $D_s \neq D_T$ , and  $\mathcal{T}_s = \mathcal{T}_T$ . In our sEMG-based gesture recognition problem, the source and target task are the same. However, the data distribution between the source and the target domain might be different i.e.,  $D_s \neq D_T$  due to factors described in section I.

To mitigate these distribution shifts on the sEMG-based gesture recognition problem, we apply the transfer learning to our proposed lightweight All-ConvNet [27] and termed it as All-ConvNet+TL. In our setting, All-ConvNet+TL has a set of shared parameters  $\theta_s$  (e.g., all the convolutional layers in All-ConvNet) and task-specific parameters for previously learned gesture recognition tasks  $\theta_0$  (e.g., the output layer of All-ConvNet for gesture recognition and its corresponding weights), and the task-specific parameters are randomly initialized for new target tasks  $\theta_n$  (e.g., gesture recognition in a new session). Considering  $\theta_0$  and  $\theta_n$  as classifiers that operate on features parameterized by  $\theta_s$ . Drawing motivation from [40], [65-66], in this work, we adopt the following approaches to learning  $\theta_n$  while taking advantage of previously learned  $\theta_s$ , which is illustrated in Fig. 1:

- (i) **Fine-tuning** – involves optimizing  $\theta_s$  and  $\theta_n$  for the new target task, while keeping  $\theta_0$  fixed (as shown in Fig.1b). To prevent large drift in  $\theta_s$ , a low learning rate is usually used. It is possible to duplicate the original network and fine-tune it for each new target task to create a set of specialized networks.
- (ii) **Feature Extraction** –  $\theta_s$  and  $\theta_0$  remain fixed and unchanged, while the outputs of one or more layers are used as features for the new target task in training  $\theta_n$  (as shown in Fig. 1c).

The most popular methodology for transfer learning is to duplicate the pre-trained network (i.e., initialize from pre-trained weights) and fine-tune (train) the entire network for each new target task [62]. However, *fine-tuning* degrades performance on previously learned tasks from the source dataset because the shared parameters change without receiving new guidance for the source-task-specific prediction parameters. In addition, *duplicating* and *fine-tuning* all the parameters of a *pre-trained model* may also require a substantial amount of target task dataset. On the other hand, *feature extraction* usually underperforms on the target dataset because the shared parameters often fail to effectively capture some discriminative information that is crucial for the target task. To address this problem and find out a good trade-off between fine-tuning and feature extraction, we focus on answering the following research questions – *Does feature reuse take place during fine-tuning or transfer learning? And if yes, where exactly is it in the network?* We first conducted a preliminary weight (or feature) transfusion experiment, where we partially reused pre-trained weights to determine and isolate the locations where meaningful feature reuse occurs. We

perform this via a weight transfusion experiment by transferring a contiguous set of some of the pre-trained weights, randomly initializing the rest of the network, and training on the target task. We have found out that meaningful feature reuse is restricted to the lowest few layers of the network and is supported by gesture recognition accuracy and convergence speed (see Appendix A for details). Following the results of these weight (or feature) transfusion experiments, the part of the  $\theta_s$  (i.e., the first three convolutional layers of All-ConvNet) were frozen and used as a *feature extractor* and only  $\theta_s$  in the top convolutional layers were *fine-tuned*. Hence, the proposed network model allows the target task to leverage complex features learned from the source dataset and make these features more discriminative for the target task by *fine-tuning* the top convolutional layers. These transfusion results suggest we propose hybrid and more flexible approaches to transfer learning (see Appendix B).

## VI. EXPERIMENTAL SETUP

We evaluated our proposed approach on CapgMyo<sup>1</sup> dataset [26] for studying and quantifying the effects of transfer learning on the smaller, simpler, and lightweight CNN. The CapgMyo dataset was developed to provide a standard benchmark database (DB) to explore new possibilities for studying and the development of cutting-edge muscle-computer interfaces (MCIs). The CapgMyo dataset includes HD-sEMG data for 128 channels (electrodes) acquired from 23 able-bodied subjects ranging in age from 23 to 26 years, which encompasses the majority of the gestures (finger movements) encountered in activities of daily living (see in Appendix C). The sampling rate is 1000 Hz. It comprised 3 sub-databases as follows:

- (a) DB-a: contains 8 isometric and isotonic hand gestures obtained from 18 of the 23 subjects. Each gesture was performed and held for 3 to 10 s.
- (b) DB-b: contains the same gesture set as in DB-a but was obtained from 10 of the 23 subjects. Each gesture in DB-b was performed and held for approximately 3 seconds. In addition, every subject in DB-b contributed to two separate recording sessions (DB-b Session 1 and DB-b Session 2), with an inter-recording interval greater than 7 (seven) days. Inevitably, the electrodes of the array were attached at slightly different positions at subsequent recording sessions.
- (c) DB-c: contains 12 hand gestures (basic movements of the fingers) obtained from 10 of the 23 subjects. Each gesture in DB-c was performed and held for approximately 3 s as in DB-b.

From the viewpoint of MCI application scenarios, the sEMG-based gesture recognition can be categorized into three (3) scenarios:

- A. *intra-session*, in which a classifier is trained on the part of the data recorded from the subjects during one session and evaluated on another part of the data recorded from the same session,
- B. *inter-session*, in which a classifier is trained on the data recorded from the subjects in one session and tested on the data recorded in another session, and

<sup>1</sup> The dataset is made publicly accessible from the following website: [http://zju-capg.org/research\\_en\\_electro\\_capgmyo.html](http://zju-capg.org/research_en_electro_capgmyo.html).



- C. *inter-subject*, when a classifier is trained on the data from a group of subjects and tested on the data from an unseen subject.

All three sub-databases (DB-a, DB-b, and DB-c) were used for *intra-session* performance evaluation. *Inter-session* recognition of hand gestures based on sEMG typically suffers from electrode shift and positioning. Therefore, DB-b was used for *inter-session* performance evaluation. Finally, both DB-b Session 2 and DB-c were used for inter-subject performance evaluation.

For CapgMyo database, first, the power-line interferences were removed from the acquired HD-sEMG signals using a 2nd order Butterworth filter with a band-stop range between 45 and 55 Hz. Then, the HD-sEMG signals were arranged in a 2-D grid according to their electrode positioning at each sampling instant. Afterward, this grid was transformed into an instantaneous sEMG image by linearly converting the values of sEMG signals from  $mV$  to color intensity as  $[-2.5mV, 2.5mV]$  to  $[0, 255]$ . As a result, instantaneous grayscale sEMG images with a size of  $16 \times 8$  matrices were obtained. To facilitate GAP, we enhance the input HD-sEMG image size from  $16 \times 8$  to  $16 \times 16$  using horizontal mirroring. Unlike [21], this enhancement does not increase the learning parameters in the proposed All-ConvNet.

For *pre-training our proposed original model* All-ConvNet, the following configurations were adopted as in [27], the connection weights for All-ConvNet network architecture were randomly initialized using Xavier initialization scheme [52], [53] and the network was trained using Adam optimization algorithm [50]. The momentum decay and scaling decay were initialized to 0.9 and 0.999, respectively. In contrast to SGD employed in [21], [23], and [26], Adam is an adaptive learning rate algorithm, therefore it requires less tuning of the learning rate hyperparameter. For all our experiments, the learning rate of 0.001 was initialized, and smaller batches of 256 randomly chosen samples from the training dataset were fed to the network during consecutive learning iterations. We set a maximum of 100 epochs for training our All-ConvNet model. However, to prevent overfitting, we applied early stopping [54], which interrupts the training process if no improvements in validation loss are observed for 5 consecutive epochs. Batch normalization [55] was applied after the input and before each non-linearity. To further regularize the network, Dropout [56] was applied to all layers with a probability of 25%. The All-ConvNet model was trained on a workstation with an Intel(R) Xeon(R) CPU E5-2620 v3 @ 2.40GHz processor, 32 GB RAM, and an NVIDIA RTX 2080 Ti GPU. Each epoch was completed in approximately 6 s for a test on intra-session gesture recognition. The average inference time per HD-sEMG sample is  $\approx 0.0929$  ms on the above-mentioned computational set up. We have also implemented the state-of-the-art network architecture [21] for a fair comparison with our proposed lightweight sEMG-based gesture recognition algorithm. However, we have adopted the same network initialization method, optimization algorithm, and training paradigm as illustrated in [21].

## VII. EXPERIMENTAL RESULTS

The sEMG-based gesture recognition methods in the literature have usually been investigated in intra-session scenarios [21], [23], [24], [36] and [61]. However, in this work, we evaluated the performance of our proposed sEMG-based gesture recognition algorithm by leveraging lightweight All-ConvNet and transfer learning in inter-session and inter-subject scenarios in addition to intra-session gesture recognition. In the following subsections, we evaluated the performance of our proposed lightweight gesture recognition algorithms. We compared them with the state-of-the-art, more complex methods in the above-mentioned three different scenarios.

### A. Intra-Session Performance Evaluation

In this section, we evaluated the performance of sEMG-based gesture recognition in the intra-session scenario. In this scenario, usually, the data variation comes from the difference between the trials and repetitions of the hand/finger gestures performed by an individual. To mitigate this data variations or distribution time shift caused by the repetitions of the gestures in multiple trials in the same session, the state-of-the-art methods performed pre-training their proposed CNN using half of the training data from all the participated subjects (e.g., 18 in DB-a) in the data collection process. Then, the pre-trained model was fine-tuned using the training data from the target subject for the subject-specific classifier development. The major drawback of this approach [21] is that the same training data used for fine-tuning was also seen during pre-training. However, in [27], we argued that the proposed lightweight All-ConvNet trained from scratch using *random initialization* has the great ability to model these distribution shifts caused by the repetitions of hand gestures across multiple trials within the same session. In that setting, we proposed designing and developing a subject-specific individualized classifier using only the sEMG data available for an individual subject while executing a target task without *pre-training*. For example, in CapgMyo DB-a and DB-b, eight (8) isotonic and isometric hand gestures were performed by an individual subject. Each gesture was also trialed and recorded 10 times with a 1000 Hz sampling rate. Thus, an individual subject generates  $(8 \times 10 \times 1000 = 80,000)$  instantaneous sEMG images. In CapgMyo DB-c, an individual performed twelve (12) basic movements of the fingers, and hence it generates  $(12 \times 10 \times 1000 = 120,000)$  instantaneous sEMG images. For performance evaluation of the proposed subject-specific lightweight All-ConvNet, a leave-one-trial-out cross-validation was performed, in which each of the 10 trials was used in turn as the test set, and the proposed lightweight All-ConvNet was trained and validated using the remaining 9 trials. This entire paradigm of training and testing process is illustrated in Fig. 1a, which shows that only the trained model (without any feature reuse from the pre-trained model) is used for gesture recognition. It is noteworthy that, in [27], we conducted experiments only on the CapgMyo DB-a and reported and compared the results with the state-of-the-art for sEMG-based gesture recognition because the maximum number of subjects (18) participated in DB-a. However, in this work, we extended our experiments on the CapgMyo DB-b and DB-c, respectively. Table II presents the gesture recognition results for the

TABLE II. THE AVERAGE RECOGNITION ACCURACIES (%) OF 8 HAND GESTURES FOR CAPGMYO DB-A AND DB-B FOR 18 AND 10 DIFFERENT SUBJECTS RESPECTIVELY AND 12 GESTURES FOR 10 DIFFERENT SUBJECTS IN DB-C. THE NUMBERS ARE MAJORITY VOTED RESULTS USING 160 MS WINDOW (I.E., 160 FRAMES). PER-FRAME ACCURACIES ARE SHOWN IN PARENTHESIS.

Model	S-ConvNet [25]	W.Geng et. al., [21]	All-ConvNet (proposed)
CapgMyo DB-a	<b>98.36 (87.95)</b>	98.48 (86.92)	98.02 (86.73)
CapgMyo DB-b Session 1	<b>97.87 (83.57)</b>	97.04 (81.26)	97.52 (81.95)
CapgMyo DB-b Session 2	<b>97.05 (84.73)</b>	96.26 (83.21)	96.80 (83.36)
CapgMyo DB-c	<b>95.80 (81.63)</b>	96.36 (82.23)	95.76 (80.91)
#Learning Parameters	$\approx 2.09 M$	$\approx 5.63 M$	$\approx 0.46 M$
Avg-run time (s)	191.29	804.66	224.33

proposed lightweight All-ConvNet and compares them with the state-of-the-art methods.

As can be seen in Table II, the proposed lightweight All-ConvNet (with around 0.46 million learning parameters) consists of a stack of  $3 \times 3$  convolutional layers with occasional subsampling by a stride of 2. It is trained from random initialization and outperformed the state-of-the-art, more complex GengNet [21], [23], [24], [26] and [61] on the CapgMyo DB-b Session 1 and Session 2 datasets, respectively, and performs comparably to the S-ConvNet [25]. Additionally, the lightweight All-ConvNet performs very competitively or on par with the GengNet [21] and S-ConvNet [25] on the CapgMyo DB-a and CapgMyo DB-c datasets, respectively. Fig. 4 (a)-(d) presents the sEMG-based instantaneous (or per-frame) gesture recognition accuracies and their statistical significance obtained through leave-one-trial-out cross-validation for ten different test trials for each of the participating subjects in CapgMyo DB-a, DB-b, and DB-c, respectively. The highest instantaneous (or per-frame) gesture recognition accuracies were 86.73% for DB-a, 81.95% and 83.36% for DB-b (Session 1 and Session 2, respectively), and 80.91% for DB-c. Which were obtained with the proposed lightweight All-ConvNet. The high per-frame gesture recognition accuracies and low standard deviation over multiple test trials and subjects in each of the four HD-sEMG datasets mentioned above reflect the high stability of the proposed lightweight All-ConvNet.

In addition, based on a simple majority voting algorithm, we have obtained very good gesture recognition accuracies. Fig. 5 (a)-(d) presents gesture recognition accuracy with different voting windows using lightweight All-ConvNet. The average gesture recognition accuracy of 94.56% and 95.99% were achieved by a simple majority voting with 32 and 64 instantaneous images (or frames) for the above four (4) HD-sEMG datasets.

The higher gesture recognition accuracies of 98.02%, 97.52%, 96.80%, and 95.76% (as shown in Table II and Fig. 5) can be obtained by the proposed lightweight All-ConvNet and a simple majority voting over the recognition result of 160 frames for DB-a, DB-b (Session 1 and Session 2) and DB-c, respectively. These outstanding results confirm that the proposed lightweight All-ConvNet is highly effective for learning all the invariances for low-resolution instantaneous HD-sEMG image recognition and hence seem to be enough to address the problem of employing high-end resource-bounded fine-tuned pre-trained

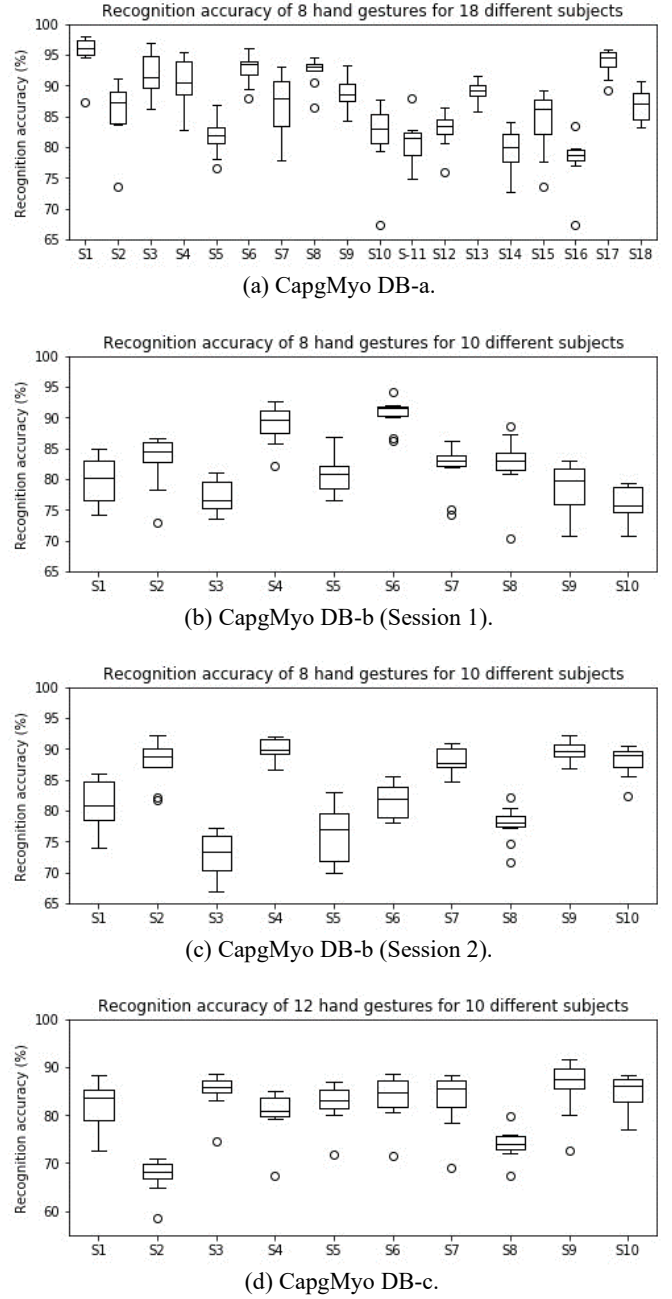


Fig 4 The per-frame gesture recognition accuracy with our proposed lightweight All-ConvNet, a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, b) and c) The gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo DB-b (Session 1) and DB-b (Session 2), respectively, and d) the gesture recognition accuracy of 12 hand gestures for 10 different subjects on CapgMyo DB-c.

networks for low-resolution instantaneous HD-sEMG image recognition.

Table II also includes average run-time for training, validation and inference for an intra-subject test. For a fair run-time comparison, each of the compared models was trained for 100 epochs on the same size of the input HD-sEMG image and early stopping [56] was applied while training all the compared models. The proposed lightweight All-ConvNet exhibits

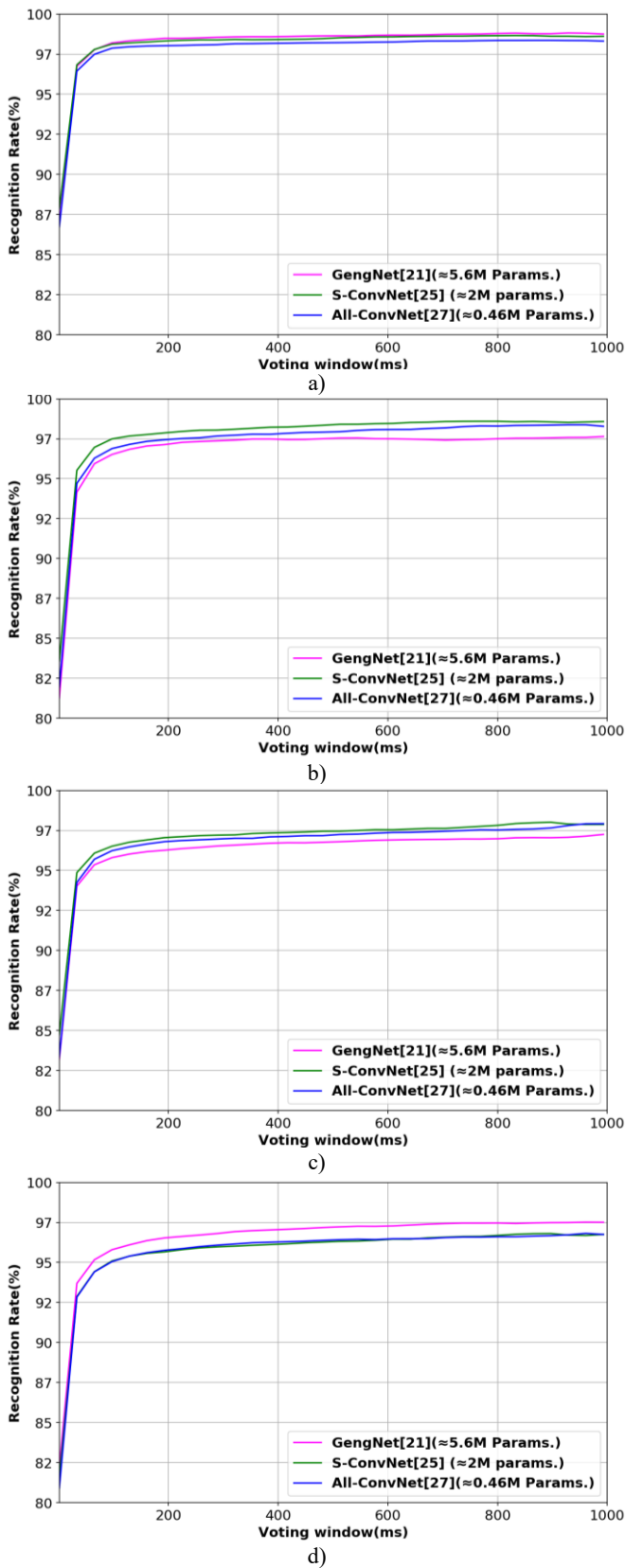


Fig 5 Surface EMG gesture recognition accuracy with different voting windows using the proposed lightweight All-ConvNet and compared with the state-of-the-art methods: a) the recognition accuracy of 8 hand gestures for 18 different subjects on CapgMyo DB-a, and the gesture recognition accuracy of 8 hand gestures for 10 different subjects on CapgMyo for b) DB-b Session 1 and c) DB-b Session 2, and d) the recognition accuracy of 12 hand gestures for 10 different subjects on DB-c.

superior run-time performance compared to the state-of-the-art methods.

### B. Inter-Session Performance Evaluation

In this section, we evaluated the performance of sEMG-based gesture recognition in the inter-session scenario. In this scenario, there is still the intra-session variability discussed in the previous section, in addition to the extent of data variability, which comes from the differences between the recording sessions. The sensor placement may have some spatial shifts and/or rotations at each recording session. These differences in sensor placement and/or rotations may cause spatial shifts in the distributions of the sEMG sensor data. To address this spatial shift problem, currently [26] and [57] provide a state-of-the-art solution in the CapgMyo dataset. Du et al. [26] proposed a multi-source extension to classical AdaBN [37] for domain adaptation. However, when dealing with multiple sources (i.e., multiple subjects), specific constraints and considerations must be imposed for each source during the model's pre-training phase [57]. Ketyko et al. [57] introduced a 2-Stage recurrent neural network (2SRNN) involving pre-training a deep stacked RNN sequence classifier on the source dataset, freezing its weights, and simultaneously training a supervised fully connected layer without a non-linear activation function on the target dataset for domain adaptation. However, ConvNet is more powerful at extracting discriminative features than RNN, even for classification tasks of long sequences [58], [59].

In addition, it is noteworthy that the domain adaptation was conducted in unsupervised and semi-supervised settings [26]. However, very low gesture recognition accuracies were reported in [26] in both inter-session and inter-subject scenarios. On the other hand, [57] performed domain adaptation in supervised settings and demonstrated state-of-the-art results on the CapgMyo dataset. Therefore, for a fair comparison with the state-of-the-art, we performed domain adaptation in a supervised manner in all the compared methods. Moreover, it might be an interesting question why we chose to compare the performance of our proposed lightweight All-ConvNet+TL with the CNN models, proposed in [21] and [26]. To the best of our knowledge, the base CNN models proposed in [21] and [26] were also adapted in [23], [24], and [61], respectively, and reported state-of-the-art results on various sEMG-based gesture recognition tasks and datasets.

Experiments conducted on inter-session and inter-subject settings; we have shown that our proposed lightweight All-ConvNet+TL leveraging transfer learning (illustrated in Section V) outperformed these above-mentioned state-of-the-art solutions. We evaluated inter-session gesture recognition for CapgMyo DBb, in which the model was trained using data recorded from the first session and evaluated using data recorded from the second session. It is worth mentioning that without transfer learning or domain adaptation, the state-of-the-art models, as well as our proposed models achieved less than or approximately 50% average gesture recognition accuracy on CapgMyo datasets in both inter-session and inter-subject scenarios. This level of recognition accuracy is not enough for a usable system (defined as <10% error [60]). Therefore, domain adaptation or transfer learning must be introduced to these (inter-session and inter-subject) settings for acceptable

performance. However, the most significant question is how much training data is required for adaptation on the target domain to obtain a stable gesture recognition accuracy. To address this question, we limited the available training data to 20% (T1), 40% (T2), 60% (T3), 80% (T4), and 100% (T5) of the total 5 trials used for domain adaptation (the remaining 5 trials are kept for validation). For fair comparison and complying with the state-of-the-art, we ran our domain adaptation for 100 epochs. Table III presents the *inter-session* average gesture recognition accuracies (%) of 8 hand gestures for 10 different subjects respectively for CapgMyo DB-b and compared with the state-of-the-art methods.

Our proposed lightweight All-ConvNet+TL leverages transfer learning to enhance inter-session gesture recognition, achieving an 11.11% improvement compared to 2SRNN [57] and a 6.43% improvement compared to GengNet [21][26] when all available 5 trials are used for adaptation (as shown in Table III, column-T5). We also compared our proposed lightweight All-ConvNet+TL with the state-of-the-art GengNet [21][26] in a data-starved condition. The proposed lightweight All-ConvNet+TL shows even more significant improvement over the state-of-the-art when a limited number of trials are available for adaptation, as seen in Table III, Columns- T1, T2, T3, and T4, respectively. For example, the proposed lightweight All-ConvNet+TL achieved a 7.94% improvement over GengNet [21][26] when only 20% of the data (i.e., 1 trial) was available for adaptation (Table III, Column- T1).

### C. Inter-Subject Performance Evaluation

In this section, we evaluated the performance of sEMG-based gesture recognition in the inter-subject scenario. In this scenario, the data variability comes from the variation in muscle physiology between different subjects. In this experiment, we evaluated the inter-subject recognition of 8 gestures using the second recording session of CapgMyo DB-b and the recognition of 12 gestures using CapgMyo DB-c. We performed a leave-one-subject-out cross-validation, in which each of the subjects was used in turn as the test subject, and a lightweight All-ConvNet was pre-trained using the data of the remaining subjects. Then, this pre-trained All-ConvNet model was deployed, and adaptation was made on the data from the odd numbers of trials of the test subjects by leveraging transfer learning or domain adaptation. Finally, the adapted model was evaluated and tested using the data from the even number of trials of the test subject. We limited the available training data to 20%, 40%, 60%, 80%, and 100% of the total 5 trials used for domain adaptation (the remaining 5 trials are kept for validation). Table IV presents the average recognition accuracies (%) of 8 and 12 hand gestures for CapgMyo DB-b and DB-c for 10 subjects, respectively.

As can be seen from Table IV, our proposed lightweight All-ConvNet+TL, by leveraging transfer learning, outperformed the state-of-the-art methods in the inter-subject scenario on both CapgMyo DB-b and CapgMyo DB-c datasets, respectively. Our proposed lightweight All-ConvNet+TL demonstrates an improvement of 5.04% and 6.17% compared to 2SRNN [57], and 3.58% and 1.85% compared to GengNet [21][26] on CapgMyo DB-b and CapgMyo DB-c datasets, respectively when all available 5 trials are used for adaptation (as shown in

TABLE III. INTER-SESSION GESTURE RECOGNITION ACCURACIES ON CAPGMYO DB-B. THE AVERAGE RECOGNITION ACCURACIES (%) OF 8 HAND GESTURES FOR 10 DIFFERENT SUBJECTS RESPECTIVELY. THE NUMBERS ARE THE MAJORITY VOTED RESULTS USING 150 MS WINDOW (I.E., 150 FRAMES).

Methods	Number of available trials for adaptation				
	T1	T2	T3	T4	T5
Du et. al. [21][26]	67.97	81.77	86.02	88.10	88.48
2SRNN [57]	-	-	-	-	83.80
All-ConvNet+TL (Proposed)	<b>75.91</b>	<b>89.61</b>	<b>92.74</b>	<b>93.46</b>	<b>94.91</b>

TABLE IV. INTER-SUBJECT GESTURE RECOGNITION ACCURACIES. THE AVERAGE RECOGNITION ACCURACIES (%) OF 8 HAND GESTURES FOR CAPGMYO DB-B AND 12 HAND GESTURES FOR CAPGMYO DB-C FOR 10 DIFFERENT SUBJECTS RESPECTIVELY. THE NUMBERS ARE THE MAJORITY VOTED RESULTS USING 150 MS WINDOW (I.E., 150 FRAMES).

Methods	CapgMyo DB-b				
	Number of available trials for adaptation				
	T1	T2	T3	T4	T5
Du et. al. [21],[26]	71.81	86.52	88.66	90.32	91.36
2SRNN [57]	-	-	-	-	89.90
All-ConvNet+TL (Proposed)	<b>75.34</b>	<b>89.42</b>	<b>92.09</b>	<b>93.83</b>	<b>94.94</b>
Methods	CapgMyo DB-c				
	Number of available trials for adaptation				
	T1	T2	T3	T4	T5
Du et. al. [21],[26]	57.40	75.98	82.51	85.98	88.02
2SRNN [57]	-	-	-	-	85.40
All-ConvNet+TL (Proposed)	<b>58.47</b>	<b>78.89</b>	<b>86.02</b>	<b>89.99</b>	<b>91.57</b>

Table IV, column-T5 for both CapgMyo DB-b and CapgMyo DB-c).

Similar to the inter-session scenario, we also compared our proposed lightweight All-ConvNet+TL in the inter-subject scenario with the state-of-the-art GengNet [21], [26] in a data-starved condition. The proposed lightweight All-ConvNet+TL exhibits improvement over the state-of-the-art on CapgMyo DB-b and CapgMyo DB-c datasets when a limited number of trials are available for adaptation, as observed in Table IV, specifically in Columns T1, T2, T3, and T4, respectively. For example, when only 20% of the data (i.e., 1 trial) was available for adaptation, the proposed lightweight All-ConvNet+TL achieved a 3.53% and 1.07% improvement over GengNet [21], [26] on CapgMyo DB-b and CapgMyo DB-c, respectively (Table IV, Column- T1).

We summarise the inter-session and inter-subject improvement results in Table V over the state-of-the-art methods. As indicated there, the performance of the proposed lightweight All-ConvNet+TL is superior in all cases. The improvement achieved by the lightweight All-ConvNet+TL leveraging transfer learning in inter-session and inter-subject scenarios, exceeds those obtained through alternative state-of-the-art domain adaptation approaches.

Finally, we evaluate the performance of our proposed lightweight All-ConvNet+TL while freezing its maximum number of layers and use them as a feature extractor, and only the top convolutions layers are fine-tuned in the adaptation stage for inter-session and inter-subject gesture recognition. More explicitly, the first six (6) convolutional layers of the lightweight All-ConvNet+TL were frozen and used as a *feature extractor*. Only the top two convolutional layers with a few parameters were fine-tuned in the adaptation stage. Therefore,

TABLE V. INTER-SESSION AND INTER-SUBJECT IMPROVEMENT (%) RESULTS OBTAINED BY THE PROPOSED LIGHTWEIGHT ALL-CONVNET+TL LEVERAGING TRANSFER LEARNING.

Methods	Inter-session improvement		Inter-subject improvement	
	DB-b	DB-b	DB-b	DB-c
Du et. al. [21][26]	6.43		3.58	3.55
2SRNN [57]	<b>11.11</b>		<b>5.04</b>	<b>6.17</b>

TABLE VI. INTER-SESSION AND INTER-SUBJECT GESTURE RECOGNITION ACCURACIES (%) UNDER FULL FEATURE EXTRACTION SETTING.

Methods	Inter-session		Inter-subject	
	DB-b	DB-b	DB-b	DB-c
2SRNN [57]	83.80		89.90	85.40
All-ConvNet+TL (Proposed)	<b>91.93</b>		<b>91.56</b>	<b>85.56</b>

these experiments can be considered as a full feature extraction setting. The performance of these full feature extraction settings was compared with the more complex computationally expensive 2SRNN [57] method. A deep-stacked RNN classifier was also used as a feature extractor by freezing its weight in the domain adaptation stage. Table VI presents the inter-session and inter-subject average gesture recognition accuracies (%) of 8 and 12 hand gestures for CapgMyo DB-b and DB-c for 10 subjects, respectively. As can be seen from Table VI, our proposed lightweight All-ConvNet+TL clearly outperforms the 2SRNN [57] in both *inter-session* and *inter-subject* gesture recognition accuracy. These experimental results indicate that the proposed lightweight All-ConvNet+TL is very effective for discriminative feature extraction for improved gesture recognition in both inter-session and inter-subject scenarios.

## VIII. DISCUSSION

We address the problem of distribution shifts by adapting a lightweight model to new target domain tasks using a limited amount of data for sEMG-based inter-session and inter-subject gesture recognition. We propose All-ConvNet+TL leveraging lightweight All-ConvNet and transfer learning, which can be seen as a hybrid of feature extraction and fine-tuning, learning parameters that are discriminative for the new target task. We show the effectiveness of our method by conducting extensive experiments on CapgMyo and its four (4) publicly available HD-sEMG sub-datasets for three (3) different sEMG-based gesture recognition tasks, including *intra-session*, *inter-session*, and *inter-subject* scenarios. The results indicate that our proposed lightweight All-ConvNet and All-ConvNet+TL models outperform the more complex state-of-the-art models on various tasks and datasets.

In *intra-session* scenarios, the proposed lightweight All-ConvNet (size of only 0.46 M learning parameters), which consists of a network using nothing, but convolutions and subsampling outperformed the most complex state-of-the-art GengNet [21], [26] (size of 5.6M parameters) on CapgMyo DB-b (Session 1 and Session 2) dataset, respectively and performed on par with or very competitively on CapgMyo DB-a and CapgMyo DB-c, respectively. The high *intra-session* gesture recognition accuracies of 98.02%, 97.52%, 96.80%, and 95.76% were obtained by the proposed lightweight All-ConvNet using a simple majority voting over the

recognition result of 160 instantaneous images (or frames) for DB-a, DB-b (Session 1 and Session 2) and DB-c, respectively. For gesture recognition in *inter-session* and *inter-subject* scenarios, we apply transfer learning to our proposed lightweight All-ConvNet. Our proposed method All-ConvNet+TL leveraging the lightweight All-ConvNet and transfer learning outperforms the current state-of-the-art methods *by a large margin*, both when the data from *single trials* or *multiple trials* are available for fine-tuning and adaptation.

We achieved state-of-the-art performance for inter-session and inter-subject scenarios. The *inter-session* gesture recognition accuracy reached 94.1% on CapgMyo DB-b, which is approximately 11.11% and 6.43% higher than the current state-of-the-art [57] and [21][26], respectively.

In addition, the *inter-subject* gesture recognition accuracy reached 94.94% and 91.57% on CapgMyo DB-b and DB-c, respectively, which is 5.04% and 6.17% higher than [57] and 3.58% and 3.55% higher than the [21], [26] respectively. Moreover, the proposed lightweight models achieved state-of-art performance under full feature extraction settings in both inter-session and inter-subject scenarios.

These outstanding *state-of-the-art* inter-session and inter-subject gesture recognition performance achieved by the proposed lightweight All-ConvNet+TL models by leveraging transfer learning validates that the proposed method is highly effective in learning invariant and discriminative representations to overcome the distribution shift caused by inter-session and inter-subject data variability. This potentially indicates that the current state-of-the-art models are overparameterized for the sEMG-based gesture recognition problem.

Furthermore, the current most complex state-of-the-art models [21], [26], [57] are computationally expensive and require a huge memory space to store a massive number of parameters. Therefore, these models are usually unsuitable for deploying low-end, resource-constrained embedded and mobile devices for real-time MCI applications. Thanks to the proposed parameter-efficient All-ConvNet and All-ConvNet+TL, our model is much smaller and lightweight than these current state-of-the-art methods for sEMG-based gesture recognition.

Finally, the new experimental evidence of our proposed method about various sEMG-based gesture recognition tasks and its role will shed light on potential future directions for the community to move forward for more efficient lightweight model exploration.

## IX. CONCLUSION

For real-time Muscle-Computer Interfaces, the sEMG-based gesture recognition must address the *inter-session* and *inter-subject* distribution shifts. To address and overcome these distribution shifts, we investigate the effects of transfer learning and feature reuse on our proposed lightweight All-ConvNet. We discovered that the proposed lightweight All-ConvNet+TL, which leverages transfer learning in the *inter-session* and *inter-subject* scenarios outperforms the most complex state-of-the-art

domain adaptation methods by a large margin, both when the data from single trials or multiple trials are available for *adaptation*. The state-of-the-art performance proved that the proposed lightweight All-ConvNet+TL model is highly effective in learning invariant and discriminative representations for addressing distribution shifts in sEMG-based inter-session and inter-subject gesture recognition. This raises the question and provides evidence of overparameterization of the most complex current state-of-the-art models for sEMG-based gesture recognition tasks. We also find that significant feature reuse concentrated in lower layers and explored more flexible and hybrid transfer approaches, which retain transfer benefits and create new possibilities. In future work, we plan to deploy our proposed lightweight All-ConvNet and All-ConvNet+TL model for sEMG-based real-time adaptive and intuitive control of an active prosthesis.

**Appendix** to “Surface EMG-Based Inter-Session/Inter-Subject Gesture Recognition by Leveraging Lightweight All-ConvNet and Transfer Learning.”

#### A. Weight (or Feature) Transfusion Experiments

In this section, we investigate to identify locations where exactly in the network meaningful feature reuse takes place during transfer learning by conducting a weight (or feature) transfusion experiment. We initialize our proposed lightweight All-ConvNet+TL with a contiguous subset of the layers using pre-trained weights (weight transfusion), and the rest of the network randomly, and train on the target inter-session gesture recognition task. More explicitly, we initialize only up to layer L with pretrained lightweight All-ConvNet+TL weights, and layer L+1 onwards randomly; then train only layers L+1 onwards. Since, the weight transfusion process uses pre-trained weights, it can accelerate the training during fine-tuning of a network on the target task. Therefore, the learning speed was measured in terms of gesture recognition performance on various training epochs. Table VII presents the inter-session gesture recognition accuracy of a subject against various training epochs for different number of transfused weights. We show the learning speed and gesture recognition accuracy when transfusing from Conv1 (L-7, one layer) up to Conv8 (i.e., layer L-7 to layers L-full transfer). From the weight transfusion results, our proposed lightweight All-ConvNet+TL model perform quite stably over the different number of transfused weights. However, we observed that reusing the lowest layers (transfusing weights) leads to the greatest gain in learning speed and gesture recognition accuracy. For example, transfusing weights from layer L-7 (Conv1) up to layer L-5 (Conv3), we achieve  $\approx 98\%$  recognition accuracy after just 8 (eight) training epochs.

#### B. Lightweight All-ConvNet Network Trimming

These weight transfusion results (Appendix A) motivate us to explore hybrid approaches to transfer learning, thereby, we introduce network trimming which further optimizes the proposed lightweight All-ConvNet+TL by pruning the weights

TABLE VII. LEARNING (OR CONVERGENCE) SPEED USING VARIOUS TRAINING EPOCHS. TABLE SHOWS INTER-SESSION GESTURE RECOGNITION ACCURACIES (%) ON TEST SET. THE NUMBERS ARE MAJORITY VOTED RESULTS USING 150 MS WINDOW (I.E., 150 FRAMES). PER-FRAME ACCURACIES ARE SHOWN IN PARENTHESIS.

Weight transfusion (up to layers)	Training epochs					
	8	16	32	46	64	100
Full Transfer (L)	70.90 (64.56)	81.74 (67.84)	83.20 (68.35)	83.08 (68.33)	83.21 (68.47)	<b>83.60</b> (68.52)
L-1	87.42 (72.28)	88.21 (73.53)	90.14 (74.43)	90.01 (74.55)	89.85 (74.94)	<b>90.39</b> (75.13)
L-2	90.24 (76.35)	93.60 (78.17)	93.94 (79.62)	94.22 (80.08)	<b>94.50</b> (80.47)	94.18 (81.36)
L-3	95.01 (79.48)	95.96 (81.53)	96.42 (83.23)	96.71 (83.22)	96.99 (83.97)	<b>98.28</b> (84.67)
L-4	96.10 (81.87)	97.71 (82.59)	98.21 (85.10)	97.92 (86.17)	97.96 (86.37)	<b>98.59</b> (87.06)
L-5	97.96 (83.14)	98.40 (84.888)	99.12 (87.00)	99.12 (86.99)	99.28 (87.86)	<b>99.35</b> (88.30)
L-6	98.34 (82.93)	97.76 (85.48)	99.26 (87.24)	98.85 (87.56)	<b>99.27</b> (87.79)	99.25 (88.68)
L-7	98.10 (83.33)	98.74 (84.34)	98.93 (86.08)	<b>99.41</b> (87.22)	99.32 (88.04)	99.32 (88.21)

TABLE VIII. LEARNING (OR CONVERGENCE) SPEED USING VARIOUS TRAINING EPOCHS. TABLE SHOWS INTER-SESSION GESTURE RECOGNITION ACCURACIES (%) ON TEST SET. THE NUMBERS ARE MAJORITY VOTED RESULTS USING 150 MS WINDOW (I.E., 150 FRAMES). PER-FRAME ACCURACIES ARE SHOWN IN PARENTHESIS.

Model	# learning parameters	Training epochs			
		8	16	24	32
Lightweight All-ConvNet+TL (Proposed)	$\approx 0.46 M$	<b>96.00</b> (71.56)	96.60 (74.79)	97.60 (76.92)	97.69 (77.68)
Lightweight All-ConvNet-Slim (Proposed)	$\approx 0.19 M$	91.92 (68.98)	<b>96.90</b> (73.70)	<b>98.28</b> (75.98)	<b>98.50</b> (77.47)

of the network. We consider reusing pre-trained weights up to Conv3 (i.e., weights of layers L-7 to layers L-5 showed in Table VII) and the weights of the top of the lightweight All-ConvNet (i.e., from layers Conv4 (L-4) to Conv7 (L-1)) was pruned by halves to be even more lightweight and initializing these layers randomly. Finally, this new Lightweight All-ConvNet-Slim model was trained or fine-tuned on the target inter-session gesture recognition task. Table VIII presents the inter-session gesture recognition accuracy of a subject against various training epochs, which compares the performance of Lightweight All-ConvNet+TL vs Lightweight All-ConvNet-Slim model. The experimental results demonstrates that the lightweight All-ConvNet-Slim model can maintain the same or achieve higher performance with much smaller number of parameters. These results with variants of Lightweight All-ConvNet+TL model also highlight many new, rich and flexible ways to use transfer learning. The preprint version of this paper has been made publicly available in [67].

#### C. Gestures and the muscles involved in CapgMyo datasets

Tables IX and X illustrate gestures and all the muscles involved in CapgMyo DB-a, DB-b and DB-c respectively [26].

TABLE IX. GESTURES IN CAPGMYO DB-A AND DB-B (8 ISOTONIC AND ISOMETRIC HAND CONFIGURATIONS)

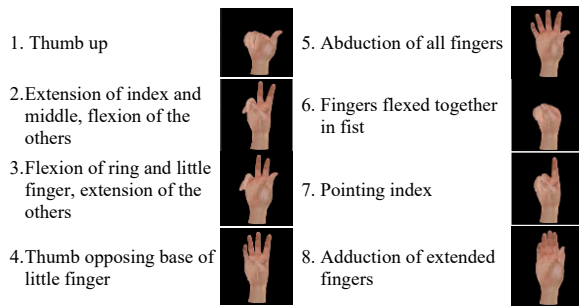
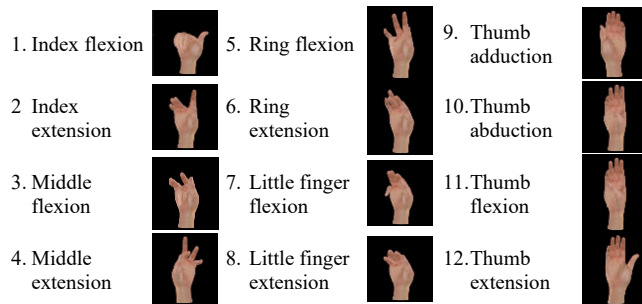


TABLE X. GESTURES IN CAPGMYO DB-C (12 BASIC MOVEMENTS OF THE FINGERS)



## REFERENCES

- [1] D. Farina *et al.*, "The extraction of neural information from the surface EMG for the control of upper-limb prostheses: merging avenues and challenges," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 22, no. 4, pp. 797–809, Jul. 2014.
- [2] G. Jang, J. Kim, J. S. Lee, and Y. Choi, "EMG-based continuous control scheme with simple classifier for electric-powered wheelchair," *IEEE Transactions on Industrial Electronics*, vol. 63, no. 6, pp. 3695–3705, 2016.
- [3] R. Jimenez-Fabian, and O. Verlinden, "Review of control algorithms for robotic ankle systems in lower-limb orthoses, prostheses, and exoskeletons," *Medical Engineering & Physics*, vol. 34, no. 4, pp. 397–408, May 2012.
- [4] Marin-Pardo, Octavio *et al.* "A Virtual Reality Muscle-Computer Interface for Neurorehabilitation in Chronic Stroke: A Pilot Study." *Sensors (Basel, Switzerland)* vol. 20,13 3754. 4 Jul. 2020.
- [5] Y. Hu, J. N. Mak, & K. Luk, "Application of surface EMG topography in low back pain rehabilitation assessment," *International IEEE/EMBS Conference on Neural Engineering*, pp. 557–560, May 2007.
- [6] D.-H. Kim, *et al.*, "Epidermal electronics," *Science*, vol. 333, pp. 838–843, 2011.
- [7] Nasri, Nadia, S. Orts-Escolano, and M. Cazorla., "An sEMG-controlled 3D game for rehabilitation therapies: real-time time hand gesture recognition using deep learning techniques," *Sensors* 20, no. 22: 6451, 2020.
- [8] T. R. Farrell & R. F. f. Weir, "A comparison of the effects of electrode implantation and targeting on pattern classification accuracy for prosthesis control," *IEEE Transactions on Biomedical Engineering*, vol. 55, no. 9, pp. 2198–2211, Sept. 2008.
- [9] M.A. Oskoei, & H. Hu. "Support vector machine-based classification scheme for myoelectric control applied to upper limb." *IEEE Transactions on Biomedical Engineering*, 55, pp.1956–1965, 2008.
- [10] Z. Lu, X. Chen, Q. Li, X. Zhang, & P. Zhou. "A hand gesture recognition framework and wearable gesture-based interaction prototype for mobile devices." *IEEE Transactions on Human-Machine Systems*, vol. 44, no. 2, pp. 293–299, 2014.
- [11] K. Li, J. Zhang, L. Wang, M. Zhang, J. Li, S. Bao, "A review of the key technologies for sEMG-based human-robot interaction systems," *Biomed. Signal Process. Control* 62 (2020), 102074, <https://doi.org/10.1016/j.bspc.2020.102074>.
- [12] E. Costanza., S. A. Inverso, R. Allen, R. and P Maes, "Intimate interfaces in action: assessing the usability and subtlety of EMG-based motionless gestures," *Conference on Human Factors in Computing Systems*, ACM, pp. 819–828, 2007.
- [13] T. S. Saponas, D. S. Tan, D. Morris, and R. Balakrishnan, "Demonstrating the feasibility of using forearm electromyography for muscle-computer interfaces," *Conference on Human Factors in Computing Systems*, ACM, pp. 515–524, 2008.
- [14] T. S. Saponas, D. S. Tan, D. Morris, D. J. Turner and J. A. Landay, "Making muscle-computer interfaces more practical," *Conference on Human Factors in Computing Systems*, pp. 851–854, ACM, 2010.
- [15] M. Atzori *et al.*, "Electromyography data for non-invasive naturally controlled robotic hand prostheses," *Scientific Data* 1, 2014.
- [16] N. Patricia, T. Tommasi. & B. Caputo, "Multi-source adaptive learning for fast control of prosthetics hand," *International Conference on Pattern Recognition*, pp. 2769–2774, 2014.
- [17] C. Amma, T. Krings, J. Ber, J. and T. Schultz, "Advancing muscle computer interfaces with high-density electromyography," *Conference on Human Factors in Computing Systems*, pp. 929–938, ACM, 2015.
- [18] A. Stango, F. Negro and D. Farina, "Spatial correlation of high-density EMG signals provides features robust to electrode number and shift in pattern recognition for myocontrol," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* Vol 23, no. 2, pp. 189–198, 2015.
- [19] M. Atzori *et al.*, "Deep learning with convolutional neural networks applied to electromyography data: A resource for the classification of movements for prosthetic hands," *Frontiers Neurobot.*, vol. 10, pp. 9–18, 2016.
- [20] X. Zhai *et al.*, "Self-recalibrating surface EMG pattern recognition for neuroprosthesis control based on convolutional neural network," *Frontiers Neurosci.*, vol. 11, pp. 379–389, 2017.
- [21] W. Geng, Y. Du, W. Jin, W. Wei, Y. Hu and J. Li, "Gesture recognition by instantaneous surface EMG images," *Scientific Reports*, Vol 15, no. 6, 36571, Nov 2016.
- [22] M. R. Islam, D. Massicotte, F. Nougrou and W. Zhu, "HOG and pairwise SVMs for neuromuscular activity recognition using instantaneous HD-sEMG images," *IEEE International New Circuits and Systems Conference (NEWCAS)*, Montreal, QC, 2018, pp. 335-339.
- [23] W.T. Wei, Y.K. Wong, Y. Du, Y. Hua, M. Kankanhallie, W.D. Geng, "A multi-stream convolutional neural network for sEMG-based gesture recognition in muscle-computer interface," *Pattern Recognit. Lett.* (2017).
- [24] Y. Hu, Y. Wong, W. Wei, Y. Du, M. Kankanhalli, W. Geng, "A novel attention-based hybrid CNN-RNN architecture for sEMG-based gesture recognition," *PLoS ONE* 13(10): e0206049. <https://doi.org/10.1371/journal.pone.0206049>, 2018.
- [25] M. R. Islam, D. Massicotte, F. Nougrou, P. Massicotte and W. -P. Zhu, "S-Convnet: a shallow convolutional neural network architecture for neuromuscular activity recognition using instantaneous high-density surface EMG images," *2020 42nd*

- Annual International Conference of the IEEE Engineering in Medicine & Biological Society (EMBC)*, 2020, pp. 744-749.
- [26] Y. Du., W. Jin, W. Wei, Y. Hu and W Geng, "Surface EMG based inter-session gesture recognition enhanced by deep domain adaptation," *Sensors*, vol. 17, no. 3, 458, 2017.
- [27] M. R. Islam, D. Massicotte and W. Zhu, "All-ConvNet: A lightweight all CNN for neuromuscular activity recognition using instantaneous high-density surface EMG images", *IEEE Int. Instrum. Meas. Technol. Conf.*, pp. 1-6, 2020.
- [28] F. Nougrou, A. Campeau-Lecours, R. Islam, D. Massicotte and B. Gosselin, "Muscle activity distribution features extracted from HDsEMG to perform forearm pattern recognition," *2018 IEEE Life Sciences Conference (LSC)*, Montreal, pp. 275-278, Oct. 2018.
- [29] Tam, M. Boukadoum, A. Campeau-Lecours and B. Gosselin, "A fully embedded adaptive real-time hand gesture classifier leveraging HD-sEMG and deep learning", *IEEE Trans. Biomed. Circuits Syst.*, vol. 14, no. 2, pp. 232-243, Apr. 2020.
- [30] F. Nougrou, A. Campeau-Lecours, D. Massicotte, M. Boukadoum, C. Gosselin, and B. Gosselin. "Pattern recognition based on HD-sEMG spatial features extraction for an efficient proportional control of a robotic arm." *Biomedical Signal Processing and Control* 53 (2019): 101550.
- [31] U. Côté-Allard et al., "Deep learning for electromyographic hand gesture signal classification using transfer learning," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 4, pp. 760-771, April 2019.
- [32] Y. Zou and L. Cheng, "A transfer learning model for gesture recognition based on the deep features extracted by CNN," *IEEE Transactions on Artificial Intelligence*, vol. 2, no. 5, pp. 447-458, Oct. 2021, doi: 10.1109/TAI.2021.3098253.
- [33] F. D. Farfan, J. C. Politti, and C. J. Felice, "Evaluation of EMG processing techniques using information theory," *Biomed. Eng. Online*, vol. 9, no. 1, pp. 1-18, 2010.
- [34] A. Krasoulis, S. Vijayakumar, and K. Nazarpour, "Multi-grip classification-based prosthesis control with two EMG-IMU sensors," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, no. 2, pp. 508-518, Feb. 2020.
- [35] Y. Taigman, M. Yang, M. Ranzato and L. Wolf, "DeepFace: closing the gap to human-level performance in face verification," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1701-1708, 23-28 June 2014.
- [36] J. Chen, Jiangcheng, B. Sheng, G. Zhang, and G. Cao., "High-density surface EMG-based gesture recognition using a 3D convolutional neural network," *Sensors* 2020, vol 20, no. 4: 1201. <https://doi.org/10.3390/s20041201>
- [37] L. Yanghao., W. Naiyan, S. Jianping, L. Jiaying and H. Xiaodi, "Revisiting batch normalization for practical domain adaptation," *arXiv:1603.04779*, 2016.
- [38] S. Ji, W. Xu, M. Yang and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221-231, Jan. 2013.
- [39] K. He, R. Girshick, and P. Dollar, "Rethinking imagenet pre-training," *IEEE International Conference on Computer Vision (ICCV)*, Seoul, 2019, pp. 4917-4926.
- [40] Z. Li and D. Hoiem, "Learning without Forgetting," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 12, pp. 2935-2947, 1 Dec. 2018.
- [41] J. Ba and R. Caruana., "Do deep nets really need to be deep?," in *Advances in neural information processing systems (NIPS)*, pages 2654-2662, 2014.
- [42] G. Hinton, O. Vinyals, and J. Dean., "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.
- [43] A. Romero, N. Ballas, S. E. Kahou, A. Chassang, C. Gatta, and Y. Bengio, "Fitnets: hints for thin deep nets," *arXiv preprint arXiv:1412.6550*, 2014.
- [44] W. Park, D. Kim, Y. Lu and M. Cho, "Relational Knowledge Distillation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3962-3971.
- [45] L. Pang, Y. Lan, J. Xu, J. Guo, and X. Cheng., "Locally smoothed neural networks," *In Proceedings of Machine Learning Research*, 77:177-191, *ACML* 2017.
- [46] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. A. Riedmiller., "Striving for simplicity: the all convolutional net," *In ICLR*, 2015. *CoRR*, abs/1412.6806
- [47] M. Lin, Q. Chen, and S. Yan, "Network in Network," *In ICLR: Conference Track*, 10 pages, 2014.
- [48] D.-A. Clevert, T. Unterthiner, and S. Hochreiter., "Fast and accurate deep network learning by exponential linear units (elus)," *arXiv preprint arXiv:1511.07289*, 2015.
- [49] K. Janocha, and W. M. Czarnecki. "On loss functions for deep neural networks in classification," *arXiv preprint arXiv:1702.05659*, 2017.
- [50] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [51] S. J. Pan and Q. Yang, "A survey on transfer learning," in *IEEE Transactions on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345-1359, Oct. 2010, doi: 10.1109/TKDE.2009.191.
- [52] X. Glorot and Y. Bengio., "Understanding the difficulty of training deep feedforward neural networks," *In AISTATS*, 2010.
- [53] K. He, X. Zhang, S. Ren, and J. Sun., "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," *In ICCV*, 2015.
- [54] R. Caruana, S. Lawrence and C. Giles, "Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping," *NIPS*, 2000.
- [55] S. Ioffe and C. Szegedy., "Batch normalization: accelerating deep network training by reducing internal covariate shift," *In ICML*, 2015.
- [56] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp.1929-1958, 2014.
- [57] I. Ketykó, F. Kovács and K. Z. Varga, "Domain adaptation for sEMG-based gesture recognition with recurrent neural networks," *arXiv:1901.06958* 2019.
- [58] K-O Cho, H-J Jang, "Comparison of different input modalities and network structures for deep learning-based seizure detection," *Sci Rep* 10, 122 (2020).
- [59] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," *In Proceedings of the 34th International Conference on Machine Learning (ICML)*, vol. 70, pp. 933-941, 2017.
- [60] E. Scheme and K. Englehart, "Electromyogram pattern recognition for control of powered upper-limb prostheses: State of the art and challenges for clinical use," *J. Rehabil. Res. Develop.*, vol. 48, no. 6, pp. 643-59, 2011, doi: 10.1682/jrrd.2010.09.0177, PMID: 21938652.
- [61] W. Wei, Q. Dai, Y. Wong, Y. Hu, M. Kankanhalli and W. Geng, "Surface-electromyography-based gesture recognition by multi-view deep learning," in *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 10, pp. 2964-2973, Oct. 2019.
- [62] M. Raghu, C. Zhang, J. Kleinberg and S. Bengio, "Transfusion: understanding transfer learning for medical imaging", *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NIPS)*, article no.: 301, Pages 3347-3357, December 2019.
- [63] M. R. Islam, D. Massicotte, F. Nougrou, P. Massicotte and W-P



- Zhu, "S-ConvNet: A shallow convolutional neural network architecture for neuromuscular activity recognition using instantaneous high-density surface EMG images," *arXiv preprint arXiv:1906.03381*, 2019.
- [64] R. N. Khushaba and K. Nazarpour, "Decoding HD-EMG Signals for Myoelectric Control - How Small Can the Analysis Window Size be?," in *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8569-8574, Oct. 2021, doi: 10.1109/LRA.2021.3111850.
- [65] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International Conference in Machine Learning (ICML)*, 2014.
- [66] Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [67] M. R. Islam, D. Massicotte, P. Massicotte and W-P Zhu, "Surface EMG-Based Inter-Session/Inter-Subject Gesture Recognition by Leveraging Lightweight All-ConvNet and Transfer Learning," *arXiv preprint, arXiv:2305.08014*, 2023.
- [68] Y. Zou, L. Cheng, L. Han, Z. Li and L. Song, "Decoding Electromyographic Signal With Multiple Labels for Hand Gesture Recognition," in *IEEE Signal Processing Letters*, vol. 30, pp. 483-487, 2023.

M. R. Islam, D. Massicotte, F. Nougrou and W. -P. Zhu, "HOG and Pairwise SVMs for Neuromuscular Activity Recognition Using Instantaneous HD-sEMG Images," *2018 16th IEEE International New Circuits and Systems Conference (NEWCAS)*, Montreal, QC, Canada, 2018, pp. 335-339, doi: 10.1109/NEWCAS.2018.8585731.

# HOG and Pairwise SVMs for Neuromuscular Activity Recognition Using Instantaneous HD-sEMG Images

*Md. Rabiul Islam<sup>1</sup>, Daniel Massicotte<sup>1</sup>, Francois Nougarou<sup>1</sup> and Wei-Ping Zhu<sup>2</sup>*

<sup>1</sup>Dept. of Electrical and Computer Engineering, Université du Québec à Trois-Rivières, QC, Canada

<sup>2</sup>Dept. of Electrical and Computer Engineering, Concordia University, Montreal, QC, Canada

**Abstract-** The concept of neuromuscular activity recognition using instantaneous high-density surface electromyography (HD-sEMG) image opens up new avenues for the development of more fluid and natural muscle-computer interfaces. The state-of-the-art methods for instantaneous HD-sEMG image recognition achieve prominent performance using a computationally intensive deep convolutional networks (ConvNet) classifier, while very low performance is reported using the conventional classifiers. However, the conventional classifiers such as Support Vector Machines (SVM) can surpass ConvNet at producing optimal classification if well-behaved feature vectors are provided. This paper studies the question of extracting distinctive feature sets, thus propose to use Histograms of Oriented Gradient (HOG) as unique features for robust neuromuscular activity recognition, adopting pairwise SVMs as the classification scheme. The experimental results proved that the HOG represents unique features inside the instantaneous HD-sEMG image and fine-tuning the hyper-parameter of the pairwise SVMs, the recognition accuracy comparable to the more complex state of the art methods can be achieved.

**Index Terms**— Neuromuscular activity recognition, HOG, HD-sEMG, Gesture recognition, SVM, Muscle-computer interface

## 1. INTRODUCTION

The precise characterization and recognition of neuromuscular activities present a great challenge [1] The high-density sEMG (HD-sEMG) based methods have been proposed in the recent years [2][3]. The HD-sEMG records myoelectric signals using two-dimensional (2D) electrode arrays that characterize the spatial distribution of myoelectric activity over the muscles that reside within the electrode pick-up area [4]. The collected HD-sEMG data are spatially correlated which enabled both temporal and spatial changes and robust against malfunction of the channels with respect to the previous counterparts [3]. However, the existing HD-sEMG based neuromuscular activity recognition methods are still depending on the windowed sEMG which demands to find an optimal window length otherwise influence the classification accuracy. To overcome this problem and develop a more fluid and natural muscle-computer interface, more recently, W. Geng et al. [4], explored the patterns inside the instantaneous sEMG images spatially composed from HD-sEMG enables neuromuscular based gesture recognition solely with the sEMG signals recorded at a specific instant. In their approach, the instantaneous values of HD-sEMG signals at each sampling instant were arranged in a 2D grid in

accordance with the electrode positioning. Afterwards, this 2D grid was converted to a grayscale sEMG image. A computational model based on deep convolutional neural networks (ConvNet) [5] has been employed for sEMG image classification. However, the potential drawback is the classification method based on ConvNet, is computationally very expensive to be practical for real-world applications for neuromuscular activity recognition. Moreover, the studies conducted in [4] reported of attaining recognition rate as low as 20% using the conventional classifiers such as support vector machines (SVM). However, the conventional classifiers such as SVM can surpass ConvNet at producing optimal classification if well-behaved feature vectors are provided [6]. However, this aspect is totally overlooked in [4]. Therefore, developing computationally efficient distinctive feature extraction and classification algorithms for instantaneous sEMG image based neuromuscular activity recognition is highly demanded.

For instantaneous sEMG image based neuromuscular activity recognition, the challenge remains open because very limited research has been done on it. This paper studies the histogram of oriented gradients (HOG) for the improved characterization of the instantaneous sEMG image. HOG is one of the state-of-the-art methods for object recognition [7]-[9]. However, this important characterization method is ignored for sEMG signal classification. This paper proposed to use a HOG based feature extraction method for instantaneous sEMG image classification. According to our best knowledge, no one performed similar studies before for sEMG signal classification.

The rest of the paper is organized as follows. Section 2 provides the computational details of the proposed feature extraction method. Section 3 describes the testing database and the experimental validation. Section 4 offers some conclusive remarks.

## 2. THE PROPOSED NEUROMUSCULAR FEATURE EXTRACTION AND CLASSIFICATION ALGORITHM

The proposed neuromuscular feature extraction and classification algorithm has three computational components: (i) preprocessing and sEMG image generation, (ii) feature extraction, and (iii) classification. A schematic diagram of the proposed muscular activity recognition method by instantaneous sEMG images are shown in Fig. 1. The sketches of hand and gestures in Fig.1 are adapted from [4].

First, the acquired HD-sEMG signals at each sampling instant were arranged in a 2-D grid according to their electrode positioning. This grid was further transformed into an instantaneous sEMG image by linearly transforming the values of sEMG signals from  $mV$  to color intensity as  $[-2.5mV, 2.5mV]$  to  $[0\ 255]$ . Thus, an instantaneous grayscale sEMG image was formed with the size of  $16 \times 8$ . The gradient image  $\nabla f(x, y)$  is obtained by convolving an estimation filters over  $x$  and  $y$  axis of the instantaneous sEMG image  $f(x, y)$ . The magnitude  $|\nabla f(x, y)|$  and orientation  $\theta(x, y)$  for each pixel of the sEMG image are computed from the gradient image  $\nabla f(x, y)$ . The sEMG image is divided into a dense grid with a spatial  $\eta \times \eta$  pixels cells. For each cell, a local 1-D histogram of gradient over all pixels in the cell are computed as features. This aggregated cell-level 1-D histogram builds the HOG feature vector for the unique representation of the instantaneous sEMG image. Finally, these HOG feature vectors are fed to a computationally effective learned pairwise SVM classifier for instantaneous gesture recognition.

Section 2.1 presents the HOG feature extraction technique for sEMG image representation and Section 2.2 presents the classification schemes respectively.

### 2.1. Histogram of Oriented Gradients (HOG) Feature Extraction

After generating the instantaneous sEMG image by linearly transforming the values of sEMG signals to color intensity as mentioned above, the crucial task is to extract distinctive features to represent the instantaneous sEMG image for robust classification of the performed hand gesture. However, the main research question is what makes the different gestures distinctive performed by the same or different subjects? For example, the hand gestures explained in Section 3 and shown in Table I can be differentiated by their shape and orientation features. The color might not be a reliable feature because the portrayed hand gestures have the same color. Therefore, any method that can precisely describe the shape and orientation information will solve the problem. Nevertheless, the problem in our hand is even more challenging because the instantaneous sEMG image is formed by linearly transforming the values of sEMG signals from  $mV$  to color intensity which reflects the intensity distributions of the performed hand gestures. The different hand gestures produce different spatial intensity distributions, thus also make the structure of the instantaneous sEMG image different. These discriminative attributes have been capitalized and used as features in this work.

Both intuitive observation and preliminary experimental results indicate that the gradient of the intensity distributions or edge directions provides the discriminative features for instantaneous sEMG image classification. HOG precisely captures this notion. Therefore, we propose to use HOG as features for instantaneous sEMG image classification. HOG features are calculated by taking orientation histograms of intensity distributions from all locations of a dense grid on a sEMG image region and combined features are used for classification. HOG features are assumed to be designed for

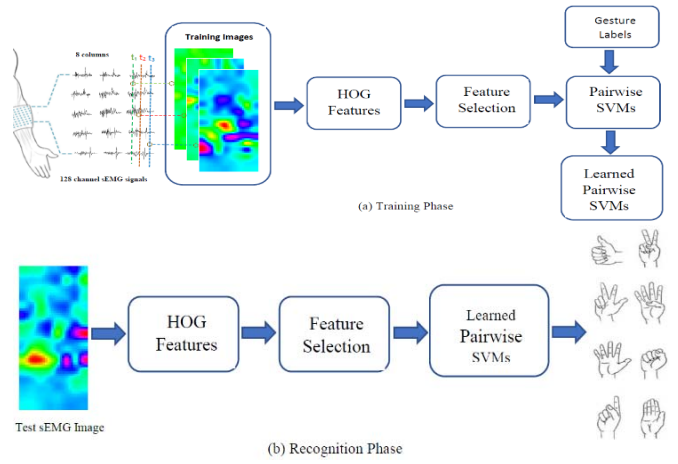


Fig. 1. Schematic illustration of the proposed muscular activity recognition by instantaneous sEMG images.

imitating the visual information processing of the brain and have robustness against local changes of position. This important property of HOG can be exploited to cope with the electrode shifting problem encountered between two different HD-sEMG recording sessions. HOG is like scale-invariant feature transform [11] in the sense that a local region is described by deriving gradient orientations from the orientation histogram.

Consider the gradient estimation filters  $h_x = [-1, 0, 1]$ , and  $h_y = [-1, 0, 1]^T$ . The gradient information of an instantaneous sEMG image can be obtained by

$$\nabla f(x, y) = \begin{bmatrix} \frac{\partial f}{\partial x} & \frac{\partial f}{\partial y} \end{bmatrix}^T = \begin{bmatrix} f(x, y) * h_x \\ f(x, y) * h_y \end{bmatrix} \quad (1)$$

where,  $*$  denotes an operation of a 1-dimensional (1-D) convolution. The  $x$  and  $y$  stand for height and width of the instantaneous sEMG image. The magnitude of a pixel is calculated by

$$|\nabla f(x, y)| = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2} \quad (2)$$

and the orientation of a pixel is calculated by

$$\theta(x, y) = \tan^{-1}\left(\frac{\partial f}{\partial x} / \frac{\partial f}{\partial y}\right) \quad (3)$$

These magnitude  $|\nabla f(x, y)|$  and orientation  $\theta(x, y)$  at each pixel are then used for calculating HOG.

The main intuition behind HOG feature extraction is that, while individual  $|\nabla f(x, y)|$  and  $\theta(x, y)$  are highly variable and subject to significant variations across nearby  $(x, y)$  locations, even for the sEMG images generated by the same hand gesture, the cumulative statistics of the spatial distribution of the gradient orientation and magnitudes over small region of the sEMG images derived from the same gesture provide quite robust descriptors of the instantaneous sEMG image.

To compute orientation histograms, the obtained instantaneous sEMG image gradient is divided into  $8 \times 4 = 32$  non-overlapping rectangular cells, and each cell is of size  $\eta \times \eta$  pixels ( $\eta = 2$ ). Four  $\eta \times \eta$  neighboring cells form a block of size  $\varsigma \times \varsigma$  ( $\varsigma = 2$ ). A schematic diagram of HOG extraction process is illustrated in Fig. 2. There are total  $v\varsigma \times$

$h\zeta = 21$ , overlapping blocks are formed over an instantaneous sEMG image (where  $v\zeta = 7$  and  $h\zeta = 3$ , denotes the number of vertical and horizontal block respectively). In each  $\eta \times \eta$  cell, the orientation histogram has  $\beta$  bins ( $\beta = 7$ ), which correspond to orientations  $i \times \pi/\beta$ , where  $i = 0, 1, \dots, \beta$ . Thus, each of the block contains  $\zeta \times \zeta \times \beta = 28$  dimensional HOG feature vectors and each instantaneous sEMG image contains  $v\zeta \times h\zeta \times (\zeta \times \zeta \times \beta) = 588$  dimensional HOG feature vectors.

This 588-dimensional HOG feature vector is used to represent the instantaneous sEMG image. It is noteworthy that  $\eta$ ,  $\zeta$  and  $\beta$  are parameters and selecting values of these parameter tradeoff with the overall instantaneous sEMG image classification performance. Therefore, it is significant to select the optimum values of these parameters for extracting most discriminant HOG features.

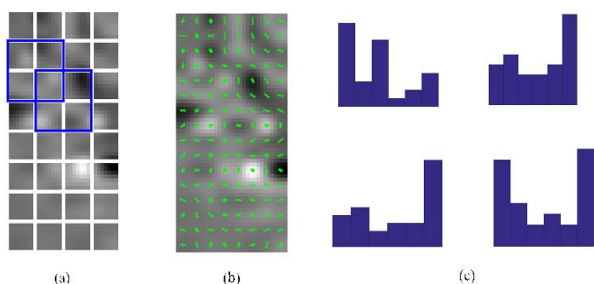


Fig. 2. HOG extraction process (a) An instantaneous sEMG image is partitioned by non-overlapping cells and overlapping blocks (each block has  $(2 \times 2)$  four cells). (b) Gradients information are overlaid over an instantaneous sEMG image (c) HOG in each block. The horizontal axis represents angle information and the vertical axis bears weighted histogram.

Now, we calculate the 28-D HOG feature vector from a block of  $\zeta \times \zeta$  cells. Consider  $|\nabla f(x, y)|$  and  $\theta(x, y)$  in one block as shown in Fig. 2(a) and 2(b). In Fig. 2(b), the orientation of the arrow represents  $\theta(x, y)$  and the length of the arrow stands for  $|\nabla f(x, y)|$ . In the experiments, the gradient orientation is transformed from  $-\pi \leq \theta \leq \pi$  to  $0 \leq \theta \leq \pi$  and then evenly quantized into  $\beta$  bins. The HOG feature vector  $h_1 \in \mathbb{R}^\beta$  of the first cell (top left in Fig. 2(a)) can be calculated by voting

$$h_1(i) \leftarrow h_1(i) + |\nabla f \theta_i(x, y)|, \quad i = 1, \dots, \beta \quad (4)$$

where  $|\nabla f \theta_i(x, y)|$  indicates the magnitude from the gradient and  $\theta_i$  is the quantized orientation. In the same way as  $h_1$ , the three-feature vectors ( $h_2, h_3$  and  $h_4$ ) can be generated from three other cells of a same block. By combining these feature vectors, the HOG feature vectors of a block turn into  $h = [h_1^T, h_2^T, h_3^T, h_4^T]^T \in \mathbb{R}^{\beta \times 4}$ .

It is to be noted that the equation (4) is a simplified form. However, in our implementation, the trilinear interpolation is used to calculate the HOG features [12]. The trilinear interpolation smoothly distributes the gradient to  $\zeta \times \zeta$  cells of a block to reduce the aliasing effect caused by the pixels near to the cell boundaries. This technique can also be robust against small distortions between sEMG images derived from the same gesture.

Moreover, the gradient strengths vary over an instantaneous sEMG image owing to local variations. Therefore, the overlapped blocks on sEMG image are normalized individually so that each scaler cell-response contributes several components to final HOG feature vector. The normalization is performed by

$$h = h / \sqrt{\|h\|_2^2 + \epsilon^2} \quad (5)$$

where,  $\epsilon$  is a small normalization constant used to avoid divided by zero [12]. This normalized HOG representation is used for instantaneous sEMG image classification.









## 2.2. Pairwise SVM Classifier

After the HOG feature extraction for representing an instantaneous sEMG image, the most important task is to employ a computationally effective classifier which has the high generalization ability for solving a multi-class classification problem. SVM [13][15] is essentially a binary classifier, however, multi-class classification problem is solved by training several binary SVM classifiers and an optimal global decision function is obtained by fusing the outputs of each of these binary classifiers. In addition, the decision function of SVM's is fully determined by the number of support vectors (SVs) which is substantially lower than the actual number of samples used in training, makes SVM computationally very efficient. Moreover, SVM trained on HOG features has become a popular method for across many visual perception tasks due to the performance and robust theory [14]. Why do SVM's trained on HOG features perform so well is still an open research issue in the literature. However, it is pointed out in [14] that preserving second-order statistics and locality of interactions are fundamental to achieve good performance. All these motivated us to use and train pairwise SVM's classifiers on HOG features extracted from the instantaneous sEMG image.

## 3. EXPERIMENTS

We tested our feature characterization method on CapgMyo data sets [10] (this database is made available from following website <http://zju-capg.org/myo/data/index.html>). This dataset was developed for providing a standard benchmark database (DB) to explore new possibilities for studying next-generation muscle-computer interfaces (MCIs). Table I illustrates gesture in DB-a and DB-b. The CapgMyo database comprises 3 sub-databases (referred as DB-a, DB-b and DB-c). However, as followed by the [4], DB-a has been used in our preliminary experiments to evaluate the performance of our proposed methods. In DB-a, 8 isotonic and isometric hand gestures were obtained from 18 of the 23 subjects and each gesture was also recorded for 10 times. For each subject, the recorded HD-sEMG data is filtered, sampled and the instantaneous sEMG image is generated using the method mentioned in Section 3. More explicitly, 8 different hand gestures are performed by every subject and each hand gestures are recorded for 10 times with a 1000 Hz sampling rate, which in total generates  $(8 \times 10 \times 1000 = 80000)$  instantaneous sEMG images. Then, our HOG-based proposed feature extraction technique elaborated in Section 2.1 is applied to each of the instantaneous sEMG images. Thus, an

Table I. Gestures in DB-a and DB-b (8 isotonic and isometric hand configurations) [10]

Label	Description	Gesture	Label	Description	Gesture
1	Thumb up		5	Abduction of all fingers	
2	Extension of index and middle flexion of others		6	Finger flexed together in fist	
3	Flexion of ring and little finger, extension of the others		7	Pointing index	
4	Thumb opposing base of little finger		8	Adduction of extended fingers	

$80000 \times M$  dimension HOG feature vectors are obtained. The each of the HOG feature vectors dimension  $M$  depend on the different HOG parameters such as  $\eta$ ,  $\zeta$  and  $\beta$ . However, considering the low resolution instantaneous sEMG image and based on our preliminary experiments, we select  $\eta = 2$ ,  $\zeta = 2$  and  $\beta = 7$  respectively. Hence, we obtained  $v\zeta \times h\zeta \times (\zeta \times \zeta \times \beta) = 588$  dimension HOG feature vectors of an instantaneous sEMG image.

Now, for every subject in DB-a, a pairwise SVMs classifier is trained to predict the desired hand gestures for each incoming sEMG images. The pairwise SVMs framework is based on LIBSVM, a library for support vector machines [16]. To conduct the above-mentioned gesture classification task, the obtained  $80000 \times M$  dimension HOG feature vectors are randomly divided into three subsets such as training, validation and testing set. In this preliminary investigation, 50% of the HOG feature vectors from the entire feature set are randomly selected and used as a training set. In the same way, the remaining 50% of the HOG feature vectors are divided into validation and testing set. The validation set is used for model/kernel and parameter selection for pairwise SVMs. Due to computationally effective and reducing searching space for parameter selection, the RBF kernel  $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$ ,  $\gamma > 0$  is used to train the obtained HOG feature set. There are two parameters for an RBF kernel which is a cost parameter ( $C$ ) and kernel parameter  $\gamma$ . It is not known in advance which  $C$  and  $\gamma$  are the best for a given problem. Therefore, the parameter selection is performed. We used a grid search along with this  $v$ -fold ( $v = 3$ ) cross-validation scheme to find the optimum  $(C, \gamma)$  on the validation set. It is recommended in [17] to use the exponentially growing sequences of  $C$  and  $\gamma$  to identify the good parameters. Hence, we use  $C = [2^5, 2^4, 2^3, \dots, 2^{-1}]$  and  $\gamma = [2^{-6}, 2^{-5}, 2^{-4}, \dots, 2^2]$ . Therefore, we examined with  $7 \times 9 = 63$  combinations of  $(C, \gamma)$  pairs. Then, the whole training feature set is trained using the pair of  $(C, \gamma)$  that achieves the best cross-validation accuracy. Finally, this trained classifier is used to predict the test feature set.

Confusion matrix generated from the predicted classification results were used as a performance indicator. The correctly classified (%) gesture classes are listed along the diagonal

line of the Confusion matrix as presented in Table II. The average classification accuracy of the proposed methods is 86.63% which is comparable to the state of the art methods. Using instantaneous values of HD-sEMG and SVM classifier, the average classification accuracy as low as 20% was reported in [4]. However, the average classification accuracy increased to 86.63% using proposed HOG and optimized parameter of pairwise SVMs. In addition, the recall or true positive rate (TPR) and the precision or the positive predictive value (PPV) [18] of each gesture classes are also computed and mentioned in Table III. The 86.62% average precision and recall of each class also indicate the potentiality of the HOG and pairwise SVMs for neuromuscular activity recognition. Finally, the experimental results demonstrate that: (i) HOG are effective features for unique representations of instantaneous HD-sEMG images (ii) Provided discriminant features and fine-tuning the hyper-parameter of the conventional classifiers such as pairwise SVMs, the state of the art recognition rate can be achieved for muscular activity recognition based on instantaneous HD-sEMG

Table II. Confusion Matrix of the Proposed Neuromuscular Activity Recognition Method.

CL01	<b>87.35</b>	0.39	1.05	0.04	0.08	7.85	3.24	0.00	
CL02	0.40	<b>86.03</b>	4.72	4.44	2.06	0.61	0.12	1.61	
CL03	1.38	3.88	<b>89.24</b>	1.74	1.29	0.44	0.89	1.13	
CL04	0.04	3.51	1.24	<b>82.84</b>	4.51	0.60	0.40	6.86	
CL05	0.04	1.76	1.12	4.31	<b>89.99</b>	0.28	0.08	2.43	
CL06	8.77	0.57	0.89	0.53	0.81	<b>83.76</b>	4.14	0.53	
CL07	2.27	0.20	1.34	0.20	0.24	4.05	<b>89.02</b>	2.67	
CL08	0.04	1.93	1.26	6.52	2.63	0.51	2.32	<b>84.79</b>	
		CL01	CL02	CL03	CL04	CL05	CL06	CL07	CL08

Table III. Precision and Recall of every gesture classes.

Class	CL01	CL02	CL03	CL04	CL05	CL06	CL07	CL08
Precision	87.52	87.44	88.38	82.32	88.57	85.07	88.66	85.03
Recall	87.35	86.03	89.24	82.84	89.99	83.76	89.02	84.79

images.

#### 4. CONCLUSIONS

In this paper, we propose to use Histogram of Oriented Gradients (HOGs) as distinctive features and pairwise SVMs for robust neuromuscular activity recognition using instantaneous HD-sEMG images. 80000 instantaneous HD-sEMG image frames for 8 different gesture of each subject from CapgMyo database were examined. The experimental results demonstrate that HOG are effective features for unique representations of instantaneous HD-sEMG images. Also, provided discriminant features and fine-tuning the hyper-parameter of the conventional classifiers such as pairwise SVMs, the state of the art recognition rate can be achieved for neuromuscular activity recognition based on instantaneous HD-sEMG images.

ACKNOWLEDGMENT – This work was supported in part by the regroupement stratégique en microsystemes du Québec (ReSMiQ) and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

## REFERENCES

- [1] Farina, D. *et al.*, "The extraction of neural information from the surface EMG for the control of upper-limb prostheses: merging avenues and challenges," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 22, 797–809, 2014.
- [2] Amma, C., Krings, T., Boer, J. & Schultz, T., "Advancing muscle-computer interfaces with high-density electromyography," *Conference on Human Factors in Computing Systems*, 929–938, ACM, 2015.
- [3] Stango, A., Negro, F. & Farina, D., "Spatial correlation of high density EMG signals provides features robust to electrode number and shift in pattern recognition for myocontrol," *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 23, 189–198, 2015.
- [4] Geng, W., Du, Y., Jin, W., Wei, W., Hu, Y., Li, J., "Gesture recognition by instantaneous surface EMG images," *Scientific Report*.2016, 6, 36571.
- [5] Krizhevsky, A., Sutskever, I. & Hinton, G. E., "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, 1097–1105, 2012.
- [6] F.J. Huang and Y. LeCun, "Large-scale Learning with SVM and Convolutional Nets for Generic Object Categorization," 2006 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 284-291, 2006.
- [7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, vol. 2, pp. 886–893, 2005.
- [8] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part Based Models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627-1645, Sept. 2010.
- [9] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical filtered motion for action recognition in crowded videos," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 3, pp. 313–323, May 2012.
- [10] Du, Y., Jin, W., Wei, W., Hu, Y., Geng, W., "Surface EMG-based inter-session gesture recognition enhanced by deep domain adaptation," *Sensors*, 17, 458, 2017.
- [11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.
- [12] N. Dalal, "Finding people in images and videos," Ph.D. thesis, INRIA Rhone-Alpes, 2006.
- [13] C.J.C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121-167, 1998.
- [14] Bristow, Hilton and Lucey, Simon, "Why do linear SVMs trained on HOG features perform so well?" in arXiv preprint arXiv:1406.2419, 2014.
- [15] C.W. Hsu and C.J. Lin, "A Comparison of Methods for Multiclass Support Vector Machines," *IEEE Trans. Neural Networks*, vol. 13, pp. 415-425, 2002.
- [16] C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology*, 2:27:1—27:27, 2011.
- [17] C. Hsu, C. C. Chang, and C. J. Lin. A practical guide to support vector classification. Technical report, 2005.
- [18] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The binormal assumption on precision-recall curves," in *Proc. 20th ICPR*, Aug. 2010, pp. 4263–4266.