A novel autonomous learning framework to enhance sEMG-based hand gesture recognition using depth information

Salih Ertug Ovur[1], Xuanyi Zhou[1,2], Wen Qi[1], Longbin Zhang[3], Yingbai Hu[4], Hang Su*[1], Giancarlo Ferrigno[1] and Elena De Momi[1]

[1] *Department of Electronics Informatics and Bioengineering, Politecnico di Milano, 20133, Milano, Italy; salihertug.ovur@mail.polimi.it*

[2] *State Key Laboratory of High Performance Complicated, Central South University, Changsha, 410083, China; zhouxuanyi@csu.edu.cn*

[3] *MoveAbility Laboratory, Department of Mechanics, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden; longbin@kth.se*

[4] *Department of Informatics, Technical University of Munich, Munich, 85748, Germany; yingbai.hu@tum.de*

Abstract

Hand gesture recognition using Surface Electromyography (sEMG) has been one of the most efficient motion analysis techniques in human-computer interaction in the last few decades. In particular, multichannel sEMG techniques have achieved stable performance in hand gesture recognition. However, the general solution of collecting and labeling large data manually leads to time-consuming implementation. A novel learning method is therefore needed to facilitate efficient data collection and preprocessing. In this paper, a novel autonomous learning framework is proposed to integrate the benefits of both depth vision and EMG signals, which automatically label the class of collected EMG data using depth information. It then utilizes a multiple layer neural network (MNN) classifier to achieve real-time recognition of the hand gestures using only the sEMG. The overall framework is demonstrated in an augmented reality application by the recognition of ten hand gestures using the Myo armband and an HTC VIVE PRO. The results show prominent

[1]*Corresponding author.
E-mail address: hang.su@polimi.it(Hang Su).

performance by introducing depth information for real-time data labeling.

Keywords: Depth vision, hands gesture recognition, machine learning, clustering, classification.

---

1. Introduction

Hand gesture is one of the most elementary and meaningful forms of human communication [1]. As a result, hand gesture recognition within Human-Computer Interaction (HCI) has received an increasing amount of attention in a wide range of applications, e.g. augmented reality [2], robot-manipulation [3, 4],rehabilitation training [5, 6] and sign language recognition [7]. Thanks to advances in computer vision technology, low-cost and innovative commercial off-the-shelf depth vision devices, such as Microsoft Kinect and Leap Motion Controller (LMC), have dramatically increased the speed of touchless interaction applications [8].

In recent years, touchless interaction in medical applications has increased since it provides more intuitive manipulation and sterile interaction compared to methods based on physical interaction with mouse and keyboard, while at the same time, offering a lower-cost solution for robot manipulation than other commercialized teleoperation devices. In [9], the touchless teleoperation of the RAVEN-II surgical robot is performed using LMC. In clinical applications, hand gestures are mostly used for the manipulation of medical image data both intraoperatively and preoperatively [10, 11, 12]. In [13], Touchless Radiology Imaging Control System (TRICS) software developed using Microsoft Kinect in order to control the image during interventional radiology procedures. Twenty-nine radiologists participated in the system evaluation survey, and the majority (69%) of them said that the proposed system could be useful for interventional radiology. In [14], a set of gestures has been introduced in order to control the projection of 2D CT scans and 3D segmentation of medical images on radiation shield during Computed Tomography (CT) by using LMC. All of these cited works show that hand gesture integration in operating rooms and telerobotic surgeries is feasible, but performance issues still exist when they are compared to clinically established methods [15, 16, 17].

Beyond the well-discussed computer vision supported hand gesture recognition algorithms, there is also a great potential in Electromyography (EMG) signals, which represents the superimposed electrical activity of muscle fibers

[18, 19]. Particularly in the field of medical robots, EMG signals are frequently employed as control signals for robotic control [20, 21, 22], which advances the intuitive interaction between human and surgical robots [23, 24, 25]. Numerous hand recognition methods using EMG signals have been proposed over the last few decades [26]. ~~In order to control a bionic manipulator, [27] a hand gesture recognition system by using the k-nearest neighbors (KNN) algorithm as a classifier for the analysis of EMG signals is required. A real-time hand gesture recognition model, using the artificial neural feedforward (ANN) network to train EMG signals, is proposed. EMG and IMU signals are combined to recognize hand gestures in order to implement hands-free navigation and free-hand writing in the air [28].~~ Nevertheless, the preprocessing of the collection and labeling of large manual data imposes a heavy work burden and results in time-consuming implementations. Moreover, pure sEMG signals are not adequate for practical applications due to the its drawbacks on low spatial resolution caused by muscle crosstalk [29]. For this purpose, in the last decade, a new kind of sensor using near-infrared light was developed to outperform sEMG approaches for detecting muscle activities [30]. More recently, myoelectrically-operated radio frequency identification (RFID) introduced to control prosthetic hands by overcoming challenges of sEMG [31]. Another solution to deal with drawbacks of sEMG came from [32], by fusing webcam and a deep learning-based EMG acquisition to control multi-functional prosthetic hands.

Lastly, contrary to the poor spatial resolution of the SEMG signals, an ultrasound sensing with a sub-millimeter spatial resolution [33] appears. To this end, a variety of studies have conducted high-precision gesture recognition with ultrasound sensing methodologies. In the [34], ten different hand gestures are classified with the more than 98% accuracy thanks to the use of the wearable ultrasonographic device by employing image processing supported neural networks. More recent research has been done by the [35]. In this research, the linear discriminant analysis (LDA) classifier compared to the support vector (SVM) classifier, where the LDA classifier obtained offline accuracy of approximately 98.83%, and the SVM classifier resulted in offline accuracy of 98.41%.

In conclusion, the accuracy of computer vision is highly dependent on the quality of the captured images, which is considerably more challenging to get in unstructured environment and dark conditions. On the other hand, most hand gesture recognition methods require a significant number of manual data preprocessing, including gesture segmentation and labeling, which

makes full automation difficult [36]. Therefore, an optimal solution is to combine depth vision and EMG in hand gesture recognition.

In this paper, a novel autonomous learning framework is proposed to automatically label the class of EMG data collected using depth vision in order to integrate the benefits of both depth vision and the EMG signal. Enhanced accuracy is tested in VR application for validation. For this purpose, firstly the depth vision acquired by HTC VIVE PRO is used to recognize hand gestures and to label EMG signals using a novel clustering algorithm. In addition, a multi-layer neural network (MNN) classifier is proposed to predict hand gestures. The results of the MNN classification are compared with the single neural network (SNN), the LDA and the SVM. Ultimately, the framework is demonstrated for the recognition of ten hand gestures using Myo Armband with the HTC VIVE PRO (3D glass). The visual output of hand posture provided to subjects using 3D glass in augmented reality mode. The novel contributions of this work include:

- A novel autonomous learning framework is presented to integrate the benefits of both depth vision and EMG signals.
- Combination of depth information and sEMG with HSOM and MNN adopted to achieve better accuracy for the designed VR application.
- A hand gesture recognition demonstration is implemented to verify the effectiveness of the proposed framework.

2. Related works

Over the last decade, a significant amount of research has been conducted using deep learning approaches to hand gesture recognition through the adoption of EMG or depth vision data [37, 38]. In order to control a bionic manipulator, a hand gesture recognition system by using the k-nearest-neighbors (KNN) algorithm as a classifier for the analysis of EMG signals is proposed in [27]. A two-channel EMG system that uses the SVM as a classifier is presented in [39] for EMG-based gesture recognition. Kinect depth sensor-based hand gesture recognition as a sign language is developed using the SVM classifier [40]. However, the SVM algorithm requires a great deal of training and can not incorporate domain knowledge [41]. A gesture detection and recognition system by decoding EMG is presented with four

time-domain features and an LDA classifier [42]. The hand gesture recognition for the real-time interface of the LDA as a classifier is designed in [43]. Although LDA excels in linear data, a small number of categorical variables are usually not practical [44].

The ANN methodology, as one of the most popular classification algorithms, has been used for various hand gesture recognition applications [45]. A number of interconnected parallel processing neurons are formed by the ANN. Each neuron receives and processes input data, and then presents the output data separately. ANN's functions can be estimated to depend on a large amount of input data. An artificial neural network classifier has been presented for depth information from the Kinect camera in hand gesture recognition [46]. An ANN-based hand gesture recognition system is proposed to process depth information applying a self co-articulated set of features [47]. A real-time hand gesture recognition model, using the ANN network to train EMG signals, is proposed in [28]. In this research, EMG and IMU signals are combined to recognize hand gestures in order to implement hands-free navigation and free-hand writing in the air. However, the aforementioned neural network methods do not realize the combined advantages of multiple neural networks. MNN primarily consists of unique neural networks trained with different initial weights and/or training data [48]. The structure of multiple neural networks is presented to improve the overall predictive performance of the system.

However, none of the above systems has achieved the combination of EMG signals and depth information. Moreover, the preprocessing and labeling of such systems is carried out manually, and therefore time-consuming. In order to solve these problems, we proposed a novel autonomous learning framework to integrate the benefits of both depth vision and EMG signal, which automatically labels the class of the collected EMG data using depth information.

Based on our previous experiences in robot-assisted minimally invasive surgery (RAMIS) [49, 3], the graphical user interface developed in our previous work [50], is improved for augmented reality (AR) in RAMIS to provide advanced visual feedback in surgical applications. In the visualization software, hand postures are overlaid in AR on real-time 3D endoscope images with scenes of painted silicone replicas of human organs.

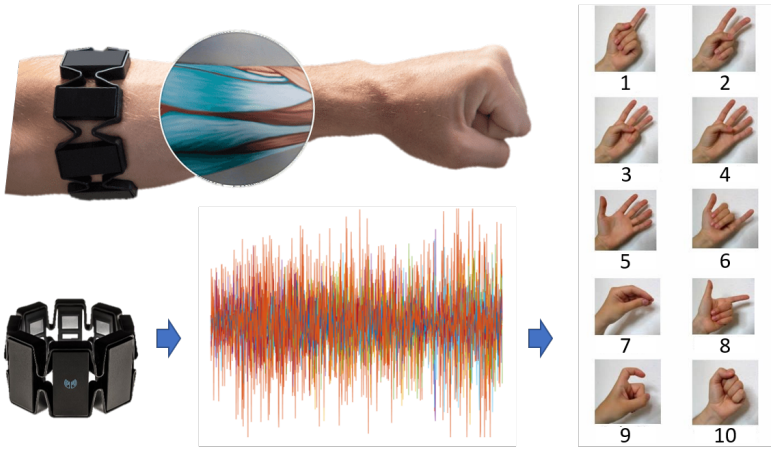


Figure 1: Myo armband in the forearm, EMG signal and Chinese number hand gestures to recognize.

3. Materials and Methods

3.1. Depth Information based Labeling

The captured depth data is used to automatically label, based on the hierarchical self-organizing map (HSOM). Three layers of the SOM model is revealed in this paper. Figure 2 displays the data stream of the self-labeling procedure. After collecting the raw depth vision data from the developed human-machine interactive system, there is a need to select useful information for accuracy enhancement. Therefore, we choose the regions of the fingertips and palm, namely $V = [V^d; V^p]^\top, V^d \in \mathbb{R}^{15}, V^p \in \mathbb{R}^{15}$. Calibration is required for the data of fingertips with the palm due to the positions and directions that are easily affected by the interference of shaking and movements. This procedure will be introduced in Section 3.2. Additionally, we adopt the wavelet denoising approach to remove the high-frequency white noise with the adaptive thresholding [51] model, as shown here:

$$G[n] = (V * \phi)[L_d] = \sum_{j=-\infty}^{\infty} V[j]\phi[L_d - i] \tag{1}$$

As the discrete mother wavelet ϕ, it can decompose the depth signals into groups of coefficients at different frequency levels. Hence, the high frequency

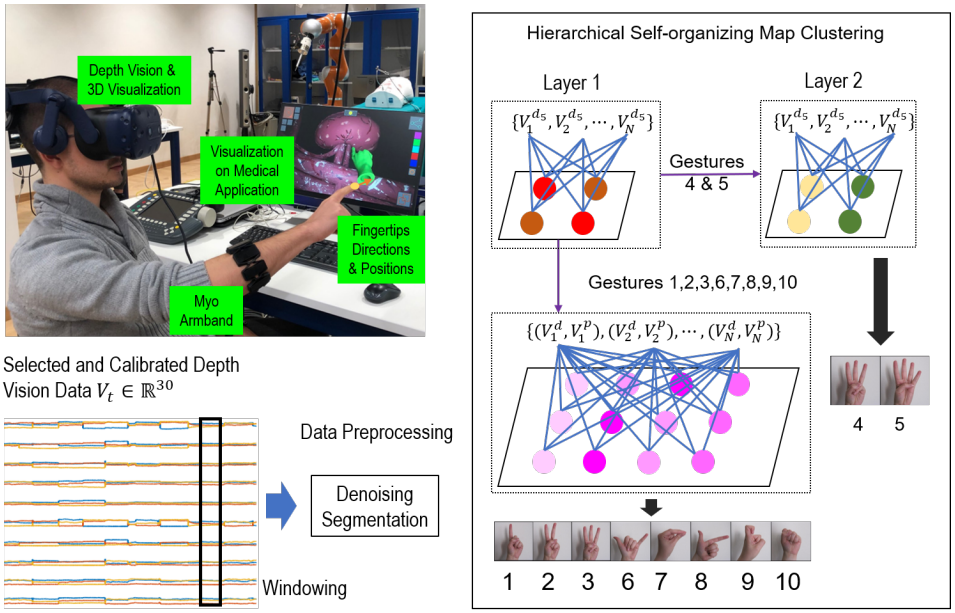


Figure 2: The flow chart of self-labeling using the three layers HSOM.

composition will be removed and the rest of the compositions are transformed by the inverse discrete wavelet transform (IDWT) method.

To label the divided depth data segments, namely $v_i, i = 1, 2, \cdots, M$, we designed a three layer HSOM clustering approach [52]. The SOM algorithm is good for large data sets, and it does not require vast amounts of memory, which results in its high performance [52]. As it is shown in Figure 2, the first step is to split the gestures into two groups, one includes gesture four and five, and the other consists of all of the rest of the gestures. To enhance the clustering accuracy, we select the data from the fifth finger, i.e., $V_5 = V_5^d, V_5^p$. The second step aims to divide gestures four and five, so we only choose the features from the first finger, i.e., $V_1 = \{V_1^d, V_1^p\}$. Due to the third step should consider separating all of the other gestures, we use all of the depth data in the last step, i.e., $V = \{V^d, V^p\}$.

In this paper, the basic SOM model is the original, stepwise recursive learning map [53]. This method will assume other real vectors $\mathbf{z}_{q,p} \in \mathbb{R}^{30}$, which represent the successively computed approximations of model $\mathbf{z}_p$. p is associated with $\mathbf{z}_p$ as the grid node in the spatial index. And a SOM

procedure is used to acquire the ordered values by

$$\mathbf{z}_{q+1,p} = \mathbf{z}_{q,p} + f_{c(q,p)} \left[v_q - \mathbf{x}_{q,p} \right] \tag{2}$$

. $f_{c(q,p)}$ is the distance function which aims to get the smallest Euclidean distance from v_q at a particular node c (winner) as follows

$$c = \underset{p}{\operatorname{argmin}} \left\{ \| v_q - \mathbf{z}_{q,p} \| \right\} \tag{3}$$

The SOM model will be modified continuously by this recursive step for obtaining the best matching model. In order to increase the computational speed, $f_{c(q,p)}$ is chosen, with its the mathematical form shown here:

$$f_{c(q,p)} = \alpha_q \cdot \exp \left[-\frac{\mathbf{D}(c,p)}{2\sigma_q^2} \right] \tag{4}$$

Where $\mathbf{D}(c,j)$ is the square of the geometric distance between the p in the grid and nodes c. α_i and σ_i are different monotonically decreasing scalar functions.

The adopted SOM model can reduce to the adaptive k-means algorithm when the neighborhood kernel value of the best matching unit (BMU) is one or zero [54]. The SOM model has proved to be a better clustering method than k-means and k-medoids approaches because it adopts the competitive learning mechanism (e.g., backpropagation with gradient descent) to label the classes [55].

3.2. Calibration of the Hand and the HTC VIVE PRO

The collected depth data from the hand should be calibrated by rotating the coordinate system from the HTC VIVE PRO RF_L to the palm RF_H, which can be described in Figure 3.

This operation can solve the problem of the impact of the palm, which results in multiple hand gestures with the same fingertip movement in different palm frames. The dynamically moving reference frame is attached to the palm frame RF_H. The transformation matrix ${}^{L}T_H$ is applied to transfer ground reference frame to the palm frame and is defined as follows:

$${}^{L}T_H = \begin{bmatrix} R_Z^{\alpha} \cdot R_Y^{\beta} \cdot R_X^{\gamma} & {}^{L}P_H \\ 0 \quad 0 \quad 0 & 1 \end{bmatrix} \tag{5}$$

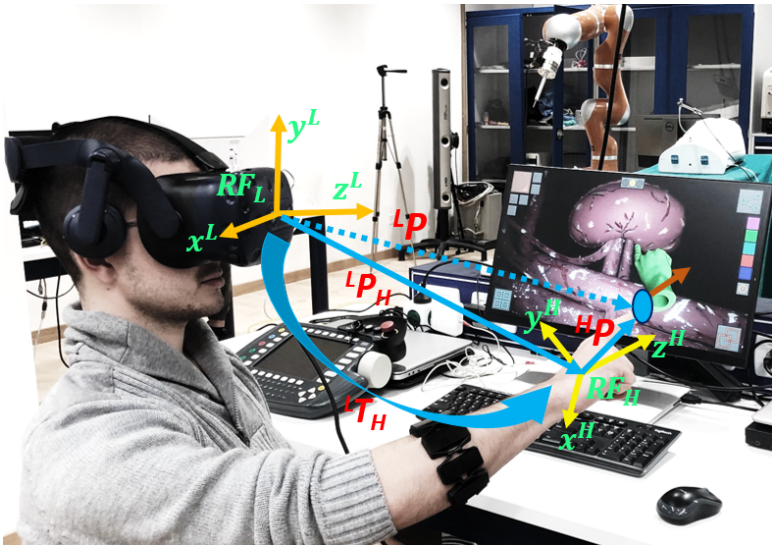


Figure 3: The coordinate transformation from HTC VIVE PRO controller to palm frame.

At time t, the angle of yaw α, pitch β, and roll γ will be rotated counter-clockwise to the palm frame. The transformation operators R_Z^α, R_Y^β and R_X^γ can be calculated by

$$R(\alpha,\beta,\gamma) = R_Z^\alpha \cdot R_Y^\beta \cdot R_X^\gamma = \begin{pmatrix} \cos\alpha\cos\beta & \cos\alpha\sin\beta\sin\gamma - \sin\alpha\cos\gamma & \cos\alpha\sin\beta\cos\gamma + \sin\alpha\sin\gamma \\ \sin\alpha\cos\beta & \sin\alpha\sin\beta\sin\gamma + \cos\alpha\cos\gamma & \sin\alpha\sin\beta\cos\gamma - \cos\alpha\sin\gamma \\ -\sin\beta & \cos\beta\sin\gamma & \cos\beta\cos\gamma \end{pmatrix} \tag{6}$$

Therefore, the coordinate system of the palm can be fixed by the designed dynamic frame between the hand and the HTC VIVE PRO. The direction and position of the five fingertips are computed according to the reference frame of the hand.

3.3. EMG-based Modeling

This work only adopts the 8D EMG signals to build the hand gesture recognition classifier. However, the EMG signal is easily affected by several types of noises, such as inherent noise in electronic equipment [56], ambient noise [57], motion artifact [58], Inherent instability of signal [59] and baseline shifts [60]. A series of signal preprocessing methods are used to remove noise and expand the raw sEMG signals, such as rectification and normalization. Recently, the multi-layers ANN approach has become the most popular

method for modeling complex classification problems and accuracy enhancement [61]. Hence, we adopt a two layer NN method to build the classifier by extracting several useful features.

3.3.1. Signal Preprocessing and Feature Extraction

The captured eight channels of EMG signals ($S \in \mathbb{R}^8$) should be processed by expanding the dimensions for acquiring more information. Figure 4 illustrates the procedure of an MNN classifier building. The Myo armband (Thalmic Labs Inc.) is worn on the user's forearm, which can provide eight dimensions (8D) EMG signals (S). To enhance the classification accuracy, we expand the raw EMG signals as follows:

$$\tilde{S} = [S; |S|; \mathcal{N}(S)]^\top = [S; |S|; \frac{S - \bar{S}}{\sigma S}]^\top \tag{7}$$

, which includes two operations of rectification $|S|$ and normalization $\mathcal{N}(S)$.

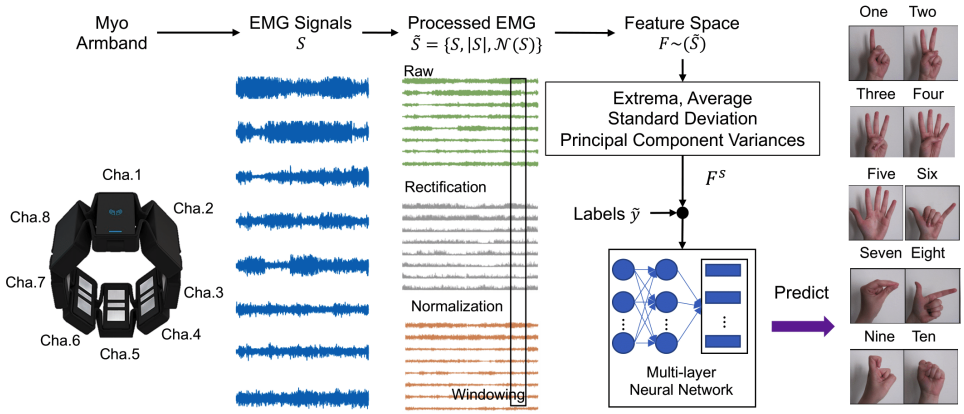


Figure 4: The EMG-based modeling procedure using MNN leanring.

As it is described in Section 3.1, the processed EMG signals will be divided into M segments as $\tilde{s}_i, i = 1, 2, \cdots, M$. Three types of features will be extracted for establishing the ensemble classifier, which is described as follows:

Extrema: meaning the maximum (ψ_α) and minimum (ψ_β). The extrema can identify the strength of the spatial region. As the minimum of the rectified EMG signal is always zero, it will not be adopted in the features space.

$$\begin{aligned} \psi_\alpha &= max(\tilde{s}_i) \\ \psi_\beta &= min(\tilde{s}_i) \end{aligned} \tag{8}$$

Standard Deviation (σ): The amount of variation can find the range and level of muscle activity.

$$\sigma = \sqrt{\frac{1}{M-1}\sum_{i=1}^{M}\left(\tilde{s}_i - \bar{s}_i\right)^2} \tag{9}$$

Where $\bar{s}_i$ denotes the average.

Average (λ): the mean of the rectified EMG signal provides the reference information of the average level of muscle activity.

$$\lambda = \bar{\tilde{s}}_i \tag{10}$$

Principal component variances (κ): it is the eigenvalues of the covariance matrix of $\tilde{s}_i$, which can be computed by principal component analysis (PCA) [62]. PCA is the most popular tool for evaluating the visualized genetic distance and relatedness between populations [63].

The sEMG segment $\tilde{s}_i$ can be dismantled into three components by singular value decomposition as

$$\tilde{s}_i = U\Lambda Q \tag{11}$$

Where U and Q are the matrices of left singular vectors and right singular vectors, respectively, the Λ is the diagonal matrix of singular values. As a column vector, the principal component variances can be calculated by the eigenvalues of the covariance matrix of $\tilde{s}_i$, which is a 1×8 vector in this paper.

In the past few decades, the artificial neural network (ANN) has become the most popular method for solving complex classification problems [64]. Although the capability of single layer ANN has proved to establish any complex model between high dimensions inputs and classes, the drawbacks of overfitting and underfitting of NN always limit the performance of the built model [65]. Therefore, we adopt the MNN to train the classifier, which consists of two feed-forward layers and a competitive layer (see Figure 5). The mapping networks in the feed-forward layer can be defined as:

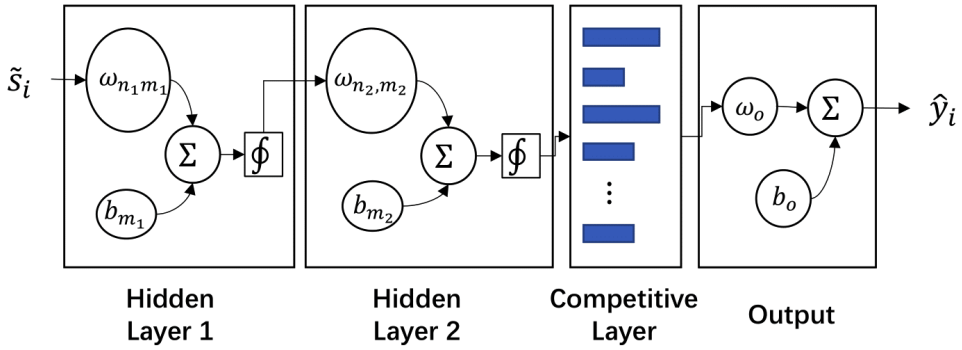


Figure 5: The schematic diagram of designed two layers feed-forward neural network (FFNN) for classification.

$$\hat{y} = b_o + \omega_o \sum_{k=1}^{K} \Phi_k \left(\sum_{n=1}^{N} \sum_{m=1}^{M} \omega_{n,m_k} \gamma_t^{m_k} + b_{m_k} \right) \tag{12}$$

Where Φ_k is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) quasi-newton activation function [66]. All of the ω and b are the weights and bias in the network, and $\gamma_t^{m_k}$ is the outputs of jth neuron. The competitive layer classifies the input vectors into a given number of classes by the similarity between vectors.

3.4. Autonomous Learning Framework

Figure 6 illustrates the proposed autonomous learning framework. In the human-machine interface, 8D sEMG signals and 30D depth vision data are synchronously captured from the HTC VIVE PRO and Myo Armband devices. Then, they will be saved into the computer. After splitting the selected depth data $V = [V^d; V^p]^\top$ into M segments, i.e., $v_i, i = 1, 2, \cdots, M$, the HSOM clustering approach can label these segments hierarchically. Even if the obtained classes $y_i^*, i = 1, 2, \cdots, M$ may not match the true gestures y, they can be considered as the ground truth. The following experiments can prove this conclusion. Similarly, the acquired sEMG signals S will be divided into M segment with the same detection length L_d, namely $\tilde{s}_i, i = 1, 2, \cdots, M$. The approaches described above will calculate the features space $(F^{\tilde{s}_i})$. The designed two layers neural network model can be trained by combining the inputs $F^{\tilde{s}_i}$ and the labels y^*, namely $\hat{y} = f(F^{\tilde{s}_i}, \omega, b)$. In real-time demonstration, this system can predict hand gestures only by adopting

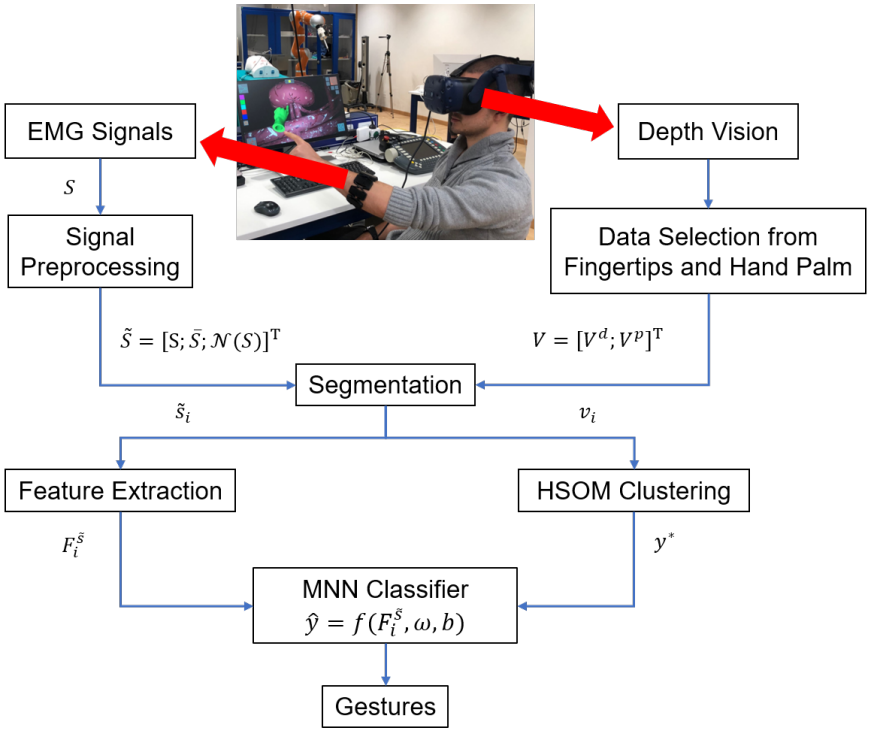


Figure 6: The pipeline of autonomous learning framework to enhance sEMG-based hand gesture recognition using depth information.

sEMG signals.

3.5. Hardware Description

In line with the proposed algorithm, multi-sensor processing and data analysis done with efficient, communicating multi-computers, as shown in Figure 7. sEMG data is collected from Myo Armband and hand is detected from depth vision feature of HTC VIVE PRO. ROS[2] with User Datagram Protocol (UDP) used to transmit data between computers. Synchronous data transfer ensured by using timestamps at the ROS messages. The first computer has an i7-4720HQ CPU 2.60 GHz processor and 8 GB RAM, and collects the EMG data from the Myo Armband and the second computer has an i7-7700HQ 2.8 GHz CPU, 8 GB GeForce 1070 GPU and 16 GB RAM, and gathers depth vision data from the HTC VIVE PRO and provides AR visualization software to the user to visualize their hands' on the surgical scene acquired by the endoscope with painted silicone replicas of organs. The sampling rate is set at 30Hz for both devices.

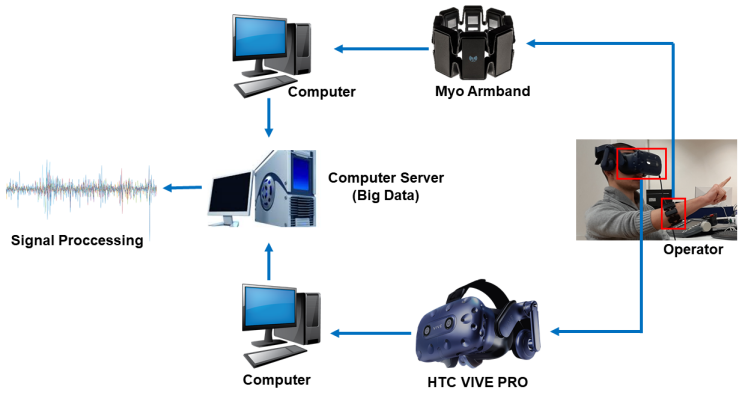


Figure 7: Overview of hardware description

Finally, data is processed, and gestures recognized by the computer acting as a server with an i9-9900K 3.6 GHz CPU, 8 GB Quadro M5000 GPU and 64 GB of RAM. The sensors of the proposed hand gesture recognition system are listed as follows:

[2]Robot Operating System, http://www.ros.org/

- The HTC VIVE PRO (HTC, New Taipei, Taiwan) is used to acquire depth vision and provides 3D AR vision on demonstrated minimally invasive surgery for the user.
- The MYO armband (Thalmic Labs, Kitchener, ON, Canada) transmits the raw EMG information over a Bluetooth Smart connection with 8 Channels (200 Hz).

4. Experimental Protocol and Results

The proposed framework consists of two main stages. The first stage is HSOM clustering for automatic labeling of collected data by using only depth vision. Labeled data (y_i^*) is used as an input in the next step to automate the system without requiring manual labeling. The second stage is classification of sEMG-based data by using the MNN classifier. Hence, a supervised learning strategy is adopted to test the accuracy of these two models. We asked ten subjects (five females and five males, between the ages of 20 and 35) to make the ten hand gestures with a fixed order from 1 to 10. They wore the Myo Gesture Control Armband is on their forearms, and they made the gestures on the HTC VIVE PRO. Each gesture took at least 3 minutes. Finally, it has $5.4e^4$ samples of both depth data and the sEMG signals because the sampling frequency is 30Hz.

For evaluating the performance of a multiple class classification problem, overall accuracy OA, F1-measure (F1-score) F_1, and receiver operating characteristic (ROC) curve are the three typical metrics. F1-measure considers both the precision P and the recall R of the test to compute the score, which can be computed by

$$F_1 = \frac{2 \times P \times R}{P + R} = \frac{2 \times \frac{TP}{TP+FP} \times \frac{TP}{TP+FN}}{\frac{TP}{TP+FP} + \frac{TP}{TP+FN}} \tag{13}$$

Where TP, FP, and FN are real positive, false positive, and false negative, respectively.

As a graphical plot, the ROC curve illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.

4.1. Clustering Analyzing

The leave-one-out strategy is adopted to calculate the average of the overall accuracy and F1-score. Table 1 displays the comparison of over-

all accuracy between HSOM with the other three popular clustering approaches, namely hierarchical k-medoids (Hk-medoids), hierarchical k-means (Hk-means), and k-means. Both Hk-medoids and Hk-means use the same three layers as the HSOM structure. The overall accuracy of HSOM is better than the others, which proves that HSOM is good at processing high dimension data streams.

Method	HSOM	Hk-medoids	Hk-means	k-means
Overall Accuracy (%)	98.42 ± 0.02	94.75 ± 0.05	85.71 ± 0.15	55.20 ± 0.22

Table 1: The comparison of overall accuracy among HSOM, Hk-medoids, Hk-means, and k-means methods.

Table 2 shows the computed F1-score of each gesture for further verification of the HSOM model's ability. Although the third gesture cannot be labeled well by HSOM, the average 90.22 is higher than that acquired by other methods.

Method	F1-measure									
	1	2	3	4	5	6	7	8	9	10
HSOM	97.43	94.99	90.22	99.90	93.44	97.01	99.90	93.89	97.99	99.90
Hk-medoids	96.38	94.25	89.89	99.90	92.78	96.26	99.90	92.54	96.90	99.34
Hk-means	89.22	92.11	86.90	90.33	91.22	90.69	94.27	93.90	89.99	92.55
k-means	67.44	69.63	78.22	74.46	69.99	73.28	76.33	75.37	78.84	75.90

Table 2: The clustering accuracy of each subject using the designed hierarchical k-medoids method

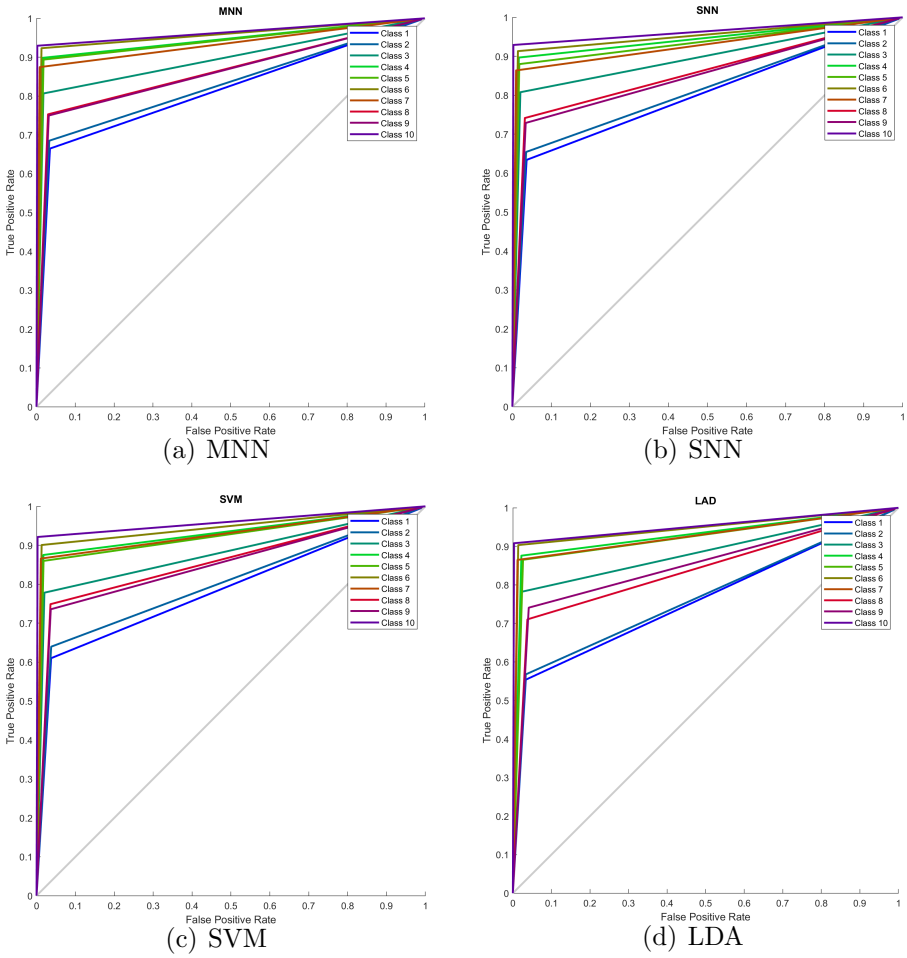
In view of the proposed depth vision guild hand gestures recognition framework should label the gestures at first, the clustering accuracy will affect the classification rate. In other words, it needs a higher clustering accuracy to obtain labels which can be regarded as the ground truth results. By comparing the designed HSOM approach with other clustering methods (i.e., Hk-medoids, Hk-means and k-means), the HSOM not only acquire the highest overall accuracy but also get the best results to label each gesture. Hence, the HSOM model is the best choice to label the depth vision segments.

4.2. MNN Evaluation

Similarly, we use the leave-one-out strategy to compare the classification performance of the MNN model with the other three machine learning approaches, i.e., single-layer neural network (SNN), SVM, and LDA classifiers.

Method	MNN	SNN	SVM	LDA
Overall Accuracy (%)	81.72 ± 0.04	80.35 ± 0.06	79.36 ± 0.05	77.86 ± 0.06

Table 3: The comparison of overall accuracy among MNN, SVM, LDA, and linear method classifiers.



(a) MNN (b) SNN

(c) SVM (d) LDA

Figure 8: The comparison of ROC curves among MNN, SNN, SVM, and LDA methods.

Table 3 displays the overall comparison accuracy among the four methods. The MNN model obtains the highest accuracy.

For further comparison of the different performance in each gesture classification, we draw the ROC curves using the four classifiers. Figure 8 shows the comparison of the ten ROC curves obtained by each method. Even if the MNN model is is comparatively better than the other three approaches, the trend of the identification for each gesture is similar. For example, the first gesture gets the worst classification result using all of the four models.

By observing the confusion matrix of the MNN classifier shown in Figure 9, high confusion rates of gestures one and nine easily detected. It is because both the first and ninth hand gesture requires the use of the index finger. As it is shown in Figure 4, the difference between these two gestures is not apparent, which will decrease the accuracy. Also, we noticed that relatively high confusion rates between Chinese gestures two and three is because of the difficulties of making gesture three without opening the little finger which was also common in similar studies [67].

Although the obtained classification accuracy of MNN classifier is higher than the other approaches, it is only over 80% which is not enough to reach a high quality for online prediction. The reason for this result is insufficient sampling and the number of classes are too much to acquire a better accuracy. However, the results of MNN (81.72%) is good enough to prove the ability of MNN method to identify these ten gestures.

5. Conclusion and Future Work

A novel autonomous learning framework is proposed in this paper for enhancing the sEMG-based hand gesture recognition. It adopts the depth information to label the ten hand gestures automatically, which are captured from the HTC VIVE PRO Controller. For robustness and reducing the interference of hand movement, a dynamically moving reference frame is designed to transfer the palm frame to the HTC VIVE PRO device. The MNN method is used to build the classifier for accuracy enhancement, which can acquire better accuracy than the other method. The proposed framework can not only be utilized for hand gesture recognition using only sEMG signals but also label the data based on depth data.

Although the proposed framework automatically achieves hand gesture recognition, further research is needed to fully implement it in RAMIS. In our future works, a 3D segmented preoperative model will be used for AR

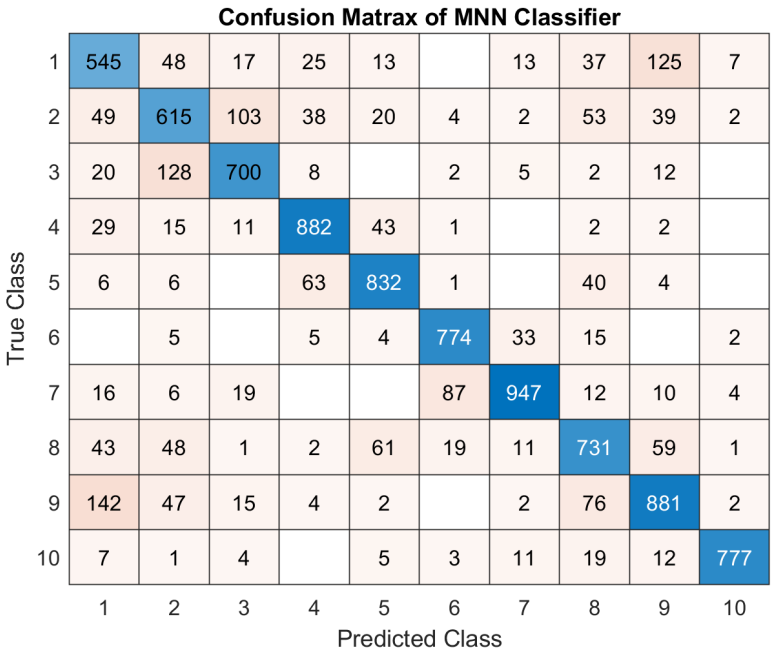


Figure 9: Confusion matrix of the MNN classifier.

navigation of surgeons, and hand gestures will manipulate it. Finally, hand gesture-based teleoperation of a surgical robot will be performed in the future.

Conflict of Interest

The authors do not have any conflict of interest for this work.

Funding

This work was supported by the European Unions Horizon 2020 research and innovation program under SMARTsurg project grant agreement No. 732515.

6. Bibliography

References

[1] J. P. Wachs, M. Kölsch, H. Stern, Y. Edan, Vision-based hand-gesture applications, Communications of the ACM 54 (2011) 60–71.

[2] Y. Pulijala, M. Ma, A. Ayoub, VR Surgery: Interactive Virtual Reality Application for Training Oral and Maxillofacial Surgeons using Oculus Rift and Leap Motion, Springer International Publishing, Cham, 2017, pp. 187–202. doi:10.1007/978-3-319-51645-5_8.

[3] H. Su, A. Mariani, S. E. Ovur, A. Menciassi, G. Ferrigno, E. De Momi, Toward teaching by demonstration for robot-assisted minimally invasive surgery, IEEE Transactions on Automation Science and Engineering (2021).

[4] X. Wu, Z. Li, Cooperative manipulation of wearable dual-arm exoskeletons using force communication between partners, IEEE Transactions on Industrial Electronics (2019).

[5] Z. Li, C. Xu, Q. Wei, C. Shi, C. Y. Su, Human-inspired control of dual-arm exoskeleton robots with force and impedance adaptation, IEEE Transactions on Systems, Man, and Cybernetics: Systems 50 (2020) 5296–5305.

[6] Z. Li, Y. Yuan, L. Luo, W. Su, K. Zhao, C. Xu, J. Huang, M. Pi, Hybrid brain/muscle signals powered wearable walking exoskeleton enhancing motor ability in climbing stairs activity, IEEE Transactions on Medical Robotics and Bionics 1 (2019) 218–227.

[7] Cao Dong, M. C. Leu, Z. Yin, American sign language alphabet recognition using microsoft kinect, 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 44–52. doi:10.1109/CVPRW.2015.7301347.

[8] Q. De Smedt, H. Wannous, J.-P. Vandeborre, Skeleton-based dynamic hand gesture recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016, pp. 1–9.

[9] F. Despinoy, N. Zemiti, G. Forestier, A. Sánchez, P. Jannin, P. Poignet, Evaluation of contactless human–machine interface for robotic surgical training, International Journal of Computer Assisted Radiology and Surgery 13 (2018) 13–24.

[10] P. Julien, L. Tommaso, A. Lounis, M. Benassarou, P. Mathieu, D. Bernot, S. Aubry, Leap motion gesture control with carestream soft-

ware in the operating room to control imaging, Surgical innovation 22 (2015).

[11] A. Mewes, B. Hensen, F. Wacker, C. Hansen, Touchless interaction with software in interventional radiology and surgery: a systematic literature review, International Journal of Computer Assisted Radiology and Surgery 12 (2017) 291–305.

[12] K. Furusawa, J. Liu, S. Tsujinaga, T. Tateyama, Y. Iwamoto, Y.-W. Chen, Robust hand gesture recognition using multimodal deep learning for touchless visualization of 3d medical images, Advances in Natural Computation, Fuzzy Systems and Knowledge Discovery, Springer International Publishing, Cham, 2020, pp. 593–600.

[13] J. Tan, C. Chao, M. Zawaideh, A. Roberts, T. Kinney, Informatics in radiology developing a touchless user interface for intraoperative image control during interventional radiology procedures, Radiographics : a review publication of the Radiological Society of North America, Inc 33 (2012).

[14] K. O'Hara, G. Gonzalez, A. Sellen, G. Penney, A. Varnavas, H. Mentis, A. Criminisi, R. Corish, M. Rouncefield, N. Dastur, T. Carrell, Touchless interaction in surgery, Communications of the ACM 57 (2014) 70–77.

[15] R. Wipfli, V. Dubois-Ferrière, S. Budry, P. J. Hoffmeyer, C. Lovis, Gesture-controlled image management for operating room: A randomized crossover study to compare interaction using gestures, mouse, and third person relaying, PloS one, 2016.

[16] S. Drouin, L. Collins, M. Kersten-Oertel, Interaction in Augmented Reality Image-Guided Surgery, 2018, pp. 99–114. doi:10.1201/9781315157702-7.

[17] S. Cronin, G. Doherty, Touchless computer interfaces in hospitals: A review, Health Informatics Journal 25 (2018) 146045821774834.

[18] A. Moin, A. Zhou, A. Rahimi, S. Benatti, A. Menon, S. Tamakloe, J. Ting, N. Yamamoto, Y. Khan, F. Burghardt, et al., An emg gesture recognition system with flexible high-density sensors and brain-inspired high-dimensional classifier, 2018 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2018, pp. 1–5.

[19] A. Rahimi, S. Benatti, P. Kanerva, L. Benini, J. M. Rabaey, Hyperdimensional biosignal processing: A case study for emg-based hand gesture recognition, 2016 IEEE International Conference on Rebooting Computing (ICRC), IEEE, 2016, pp. 1–8.

[20] P. Polygerinos, K. C. Galloway, S. Sanan, M. Herman, C. J. Walsh, Emg controlled soft robotic glove for assistance during activities of daily living, 2015 IEEE international conference on rehabilitation robotics (ICORR), IEEE, 2015, pp. 55–60.

[21] L.-Z. Liao, Y.-L. Tseng, H.-H. Chiang, W.-Y. Wang, Emg-based control scheme with svm classifier for assistive robot arm, 2018 International Automatic Control Conference (CACS), IEEE, 2018, pp. 1–5.

[22] L. Zhang, Z. Li, Y. Hu, C. Smith, E. M. G. Farewik, R. Wang, Ankle joint torque estimation using an emg-driven neuromusculoskeletal model and an artificial neural network model, IEEE Transactions on Automation Science and Engineering (2020) 1–10.

[23] A.-C. Tsai, J.-J. Luh, T.-T. Lin, A novel stft-ranking feature of multichannel emg for motion pattern recognition, Expert Systems with Applications 42 (2015) 3327–3341.

[24] H. Su, W. Qi, C. Yang, J. Sandoval, G. Ferrigno, E. De Momi, Deep neural network approach in robot tool dynamics identification for bilateral teleoperation, IEEE Robotics and Automation Letters 5 (2020) 2943–2949.

[25] W. Qi, H. Su, A. Aliverti, A smartphone-based adaptive recognition and real-time monitoring system for human activities, IEEE Transactions on Human-Machine Systems 50 (2020) 414–423.

[26] Z. Li, B. Wang, F. Sun, C. Yang, Q. Xie, W. Zhang, semg-based joint force control for an upper-limb power-assist exoskeleton robot, IEEE journal of biomedical and health informatics 18 (2013) 1043–1050.

[27] W.-T. Shi, Z.-J. Lyu, S.-T. Tang, T.-L. Chia, C.-Y. Yang, A bionic hand controlled by hand gesture recognition based on surface emg signals: A preliminary study, Biocybernetics and Biomedical Engineering 38 (2018) 126–135.

[28] J. E. E. Goh, M. L. I. Goh, J. S. Estrada, N. C. Lindog, J. C. M. Tabulog, N. E. C. Talavera, Presentation-aid armband with imu, emg sensor and bluetooth for free-hand writing and hand gesture recognition, International Journal of Computing Sciences Research 1 (2018) 65–77.

[29] W. Guo, X. Sheng, H. Liu, X. Zhu, Development of a multi-channel compact-size wireless hybrid semg/nirs sensor system for prosthetic manipulation, IEEE Sensors Journal 16 (2015) 447–456.

[30] B. Shadgan, W. D. Reid, R. Gharakhanlou, L. Stpublisher-ids, A. J. Macnab, Wireless near-infrared spectroscopy of skeletal muscle oxygenation and hemodynamics during exercise and ischemia, Journal of Spectroscopy 23 (2009) 233–241.

[31] M. S. Trachtenberg, G. Singhal, R. Kaliki, R. J. Smith, N. V. Thakor, Radio frequency identification—an innovative solution to guide dexterous prosthetic hands, 2011 annual international conference of the IEEE engineering in medicine and biology society, IEEE, 2011, pp. 3511–3514.

[32] G. Ghazaei, A. Alameer, P. Degenaar, G. Morgan, K. Nazarpour, Deep learning-based artificial vision for grasp classification in myoelectric hands, Journal of neural engineering 14 (2017) 036025.

[33] W. Xia, Y. Zhou, X. Yang, K. He, H. Liu, Toward portable hybrid surface electromyography/a-mode ultrasound sensing for human–machine interface, IEEE Sensors Journal 19 (2019) 5219–5228.

[34] J. McIntosh, A. Marzo, M. Fraser, C. Phillips, Echoflex: Hand gesture recognition using ultrasound imaging, Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 1923–1934.

[35] X. Yang, X. Sun, D. Zhou, Y. Li, H. Liu, Towards wearable a-mode ultrasound sensing for real-time finger motion recognition, IEEE Transactions on Neural Systems and Rehabilitation Engineering 26 (2018) 1199–1208.

[36] D. Holden, J. Saito, T. Komura, A deep learning framework for character motion synthesis and editing, ACM Transactions on Graphics (TOG) 35 (2016) 138.

[37] U. Côté-Allard, C. L. Fall, A. Drouin, A. Campeau-Lecours, C. Gosselin, K. Glette, F. Laviolette, B. Gosselin, Deep learning for electromyographic hand gesture signal classification using transfer learning, IEEE Transactions on Neural Systems and Rehabilitation Engineering 27 (2019) 760–771.

[38] V. Kartsch, S. Benatti, M. Mancini, M. Magno, L. Benini, Smart wearable wristband for emg based gesture recognition powered by solar energy harvester, 2018 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2018, pp. 1–5.

[39] M. Tavakoli, C. Benussi, P. A. Lopes, L. B. Osorio, A. T. de Almeida, Robust hand gesture recognition with a double channel surface emg wearable armband and svm classifier, Biomedical Signal Processing and Control 46 (2018) 121–130.

[40] S. Saha, S. Bhattacharya, A. Konar, A novel approach to gesture recognition in sign language applications using avl tree and svm, in: Progress in Intelligent Computing Techniques: Theory, Practice, and Applications, Springer, 2018, pp. 271–277.

[41] W. Zhu, Y. Wei, H. Xiao, Fault diagnosis of neural network classified signal fractal feature based on svm, Cluster Computing 22 (2019) 4249–4254.

[42] S. Pancholi, A. M. Joshi, Electromyography-based hand gesture recognition system for upper limb amputees, IEEE Sensors Letters 3 (2019) 1–4.

[43] O. Sangjun, R. Mallipeddi, M. Lee, Real time hand gesture recognition using random forest and linear discriminant analysis, Proceedings of the 3rd International Conference on Human-Agent Interaction, ACM, 2015, pp. 279–282.

[44] A. Bhardwaj, A. Gupta, P. Jain, A. Rani, J. Yadav, Classification of human emotions from eeg signals using svm and lda classifiers, 2015 2nd International Conference on Signal Processing and Integrated Networks (SPIN), IEEE, 2015, pp. 180–185.

[45] H. Liu, L. Wang, Gesture recognition for human-robot collaboration: A review, International Journal of Industrial Ergonomics 68 (2018) 355–367.

[46] S. Jadooki, D. Mohamad, T. Saba, A. S. Almazyad, A. Rehman, Fused features mining for depth-based hand gesture recognition to classify blind human communication, Neural Computing and Applications 28 (2017) 3285–3294.

[47] J. Singha, R. H. Laskar, Ann-based hand gesture recognition using self co-articulated set of features, IETE Journal of Research 61 (2015) 597–608.

[48] J. Rezaei, M. Shahbakhti, B. Bahri, A. A. Aziz, Performance prediction of hcci engines with oxygenated fuels using artificial neural networks, Applied Energy 138 (2015) 460–473.

[49] H. Su, W. Qi, Y. Hu, H. R. Karimi, G. Ferrigno, E. De Momi, An incremental learning framework for human-like redundancy optimization of anthropomorphic manipulators, IEEE Transactions on Industrial Informatics (2020).

[50] S. E. Ovur, M. Cobanaj, L. Vantadori, E. De Momi, G. Ferrigno, Surgeon training with haptic devices for computer and robot assisted surgery: An experimental study, XV Mediterranean Conference on Medical and Biological Engineering and Computing – MEDICON 2019, Springer International Publishing, Cham, 2020, pp. 1526–1535.

[51] Ç. P. Dautov, M. S. Özerdem, Wavelet transform and signal denoising using wavelet method, 2018 26th Signal Processing and Communications Applications Conference (SIU), IEEE, 2018, pp. 1–4.

[52] A. Rauber, D. Merkl, M. Dittenbach, The growing hierarchical self-organizing map: exploratory analysis of high-dimensional data, IEEE Transactions on Neural Networks 13 (2002) 1331–1341.

[53] T. Kohonen, Essentials of the self-organizing map, Neural networks 37 (2013) 52–65.

[54] W. Kurdthongmee, A hardware centric algorithm for the best matching unit searching stage of the som-based quantizer and its fpga implementation, Journal of Real-Time Image Processing 12 (2016) 71–80.

[55] S. Benedetti, S. Bucciarelli, F. Canestrari, S. Catalani, S. Mandolini, V. Marconi, A. R. Mastrogiacomo, R. Silvestri, M. C. Tagliamonte, R. Venanzini, et al., Platelet's fatty acids and differential diagnosis of major depression and bipolar disorder through the use of an unsupervised competitive-learning network algorithm (som), Open Journal of Depression 3 (2014) 52.

[56] S. N. Kale, S. V. Dudul, Intelligent noise removal from emg signal using focused time-lagged recurrent neural network, Applied Computational Intelligence and Soft Computing 2009 (2009) 1.

[57] S. Day, Important factors in surface emg measurement, Bortec Biomedical Ltd publishers (2002) 1–17.

[58] P. Mithun, P. C. Pandey, T. Sebastian, P. Mishra, V. K. Pandey, A wavelet based technique for suppression of emg noise and motion artifact in ambulatory ecg, 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, 2011, pp. 7087–7090.

[59] K. G. Anderson, D. G. Behm, Maintenance of emg activity and loss of force output with instability, The Journal of Strength & Conditioning Research 18 (2004) 637–640.

[60] D. V. Harris, W. J. Robinson, The effects of skill level on emg activity during internal and external imagery, Journal of Sport and Exercise Psychology 8 (1986) 105–111.

[61] G. Ditzler, R. Polikar, G. Rosen, Multi-layer and recursive neural networks for metagenomic classification, IEEE transactions on nanobioscience 14 (2015) 608–616.

[62] D. Staudenmann, I. Kingma, A. Daffertshofer, D. F. Stegeman, J. H. van Dieën, Improving emg-based muscle force estimation by using a high-density emg grid and principal component analysis, IEEE Transactions on Biomedical Engineering 53 (2006) 712–719.

[63] A. Gallina, S. J. Garland, J. M. Wakeling, Identification of regional activation by factorization of high-density surface emg signals: A comparison of principal component analysis and non-negative matrix factorization, Journal of Electromyography and Kinesiology 41 (2018) 116–123.

[64] H.-P. Huang, Y.-H. Liu, L.-W. Liu, C.-S. Wong, Emg classification for prehensile postures using cascaded architecture of neural networks with self-organizing maps, 2003 IEEE International Conference on Robotics and Automation (Cat. No. 03CH37422), volume 1, IEEE, 2003, pp. 1497–1502.

[65] G. Tsenov, A. Zeghbib, F. Palis, N. Shoylev, V. Mladenov, Neural networks for online classification of hand and finger movements using surface emg signals, 2006 8th Seminar on Neural Network Applications in Electrical Engineering, IEEE, 2006, pp. 167–171.

[66] J. D. Head, M. C. Zerner, A broyden—fletcher—goldfarb—shanno optimization procedure for molecular geometries, Chemical physics letters 122 (1985) 264–270.

[67] P. B. Shull, S. Jiang, Y. Zhu, X. Zhu, Hand gesture recognition and finger angle estimation via wrist-worn modified barometric pressure sensing, IEEE Transactions on Neural Systems and Rehabilitation Engineering 27 (2019) 724–732.