Sign Language Hand Gesture Recognition Using Autoencoder and Support Vector Machine Classifiers

Aida Abd El-Mohsen 1, Eman M. El-Daydamony 1, Fatma E.Z. Abou-Chadi 2

1 Department of Information Technology, Faculty of Computer and Information Science, Mansoura University, Egypt.
2 Department of Electrical Engineering, Faculty of Engineering, The British University in Egypt, Egypt.

E-mail: uogeno@gmail.com; eman.8.2000@gmail.com; F-abochadi@ieee.org

Abstract – We propose a novel 4 stage hand gesture recognition algorithm. The algorithm uses combinations of features to implement vision–based hand gesture recognition. These include Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), fingertip finder, pixel segmentation, eccentricity, elongation, and rotation. Feature fusion is adopted by combining all extracted features, which can detect hand gesture with high accuracy. In addition, score fusion method is also used to combine results achieving high accuracy rate. The Automatic sign language system requires quick and precise methods to distinguish static signs or an arrangement of signs delivered to help translate their right significance. Deep Learning algorithm and Multiclass Support Vector Machine algorithm are used to distinguish 37 hand signs of American numbers and alphabets correctly that help the Human Computer Interaction (HCI) system. The proposed system was learned and tested using 1850 image for 37 gestures of hand and reaches a 98.64% correct classification in real time environment.

Keywords - Human Computer Interaction (HCI), Histogram of Oriented Gradients (HOG), Local Binary Pattern (LBP), Deep Learning algorithm, Multiclass Support Vector Machine algorithm.

I. INTRODUCTION

Hand gesture is an effective research area, mostly used for hand gesture recognition and Human Computer Interaction (HCI). The objective is basically aiming to superior human-machine way of communicating so that it leads to the highest possible human interaction. There are around 70 million hard of hearing individuals [1] with hearing impairment and in addition hearing individuals all over the world use sign language as their mother language or first tongue. Every state has one or, in some cases, at least two sign languages, even though the various sign languages can share an equivalent linguistic root within the same way as the spoken languages. However, it is not a global language; the general characteristics in sign languages benefits users of distinct sign languages to be understood much faster than users of distinct spoken languages do.

Research of sign language recognition can be divided into two approaches. One is the sensor-based approach, which uses sensors such as Accelerometer [2], and gloves [3] that track the movement and location of the hand gesture. This approach is very expensive and inadequate for all normal users. The second approach is more suitable. It is based on computer vision. It depends on using built in mobile camera, but faces greater complexity.

Researchers have used different techniques in their work such as Convexity detection algorithm [6], two dimensional Zernike [21], Gray Level Co-Occurrence Matrix [23] etc., but it is still needed to improve the recognition accuracy.

The present work is an attempt to improve the recognition accuracy of the sign recognition system by using a combination of such techniques and comparing their performance.

The remainder of the paper is set as follows: Section II discusses related work. Section III explains material and methods that explains the proposed framework. Section IV presents results and discussion. Finally, the paper is concluded in Section V.

II. RELATED WORK

Many researchers have done plenty of work on the visual recognition of hand gesture. In [4], the authors presented an automatic system for American Sign Language detection and its translation to text. They used Hue Saturation Value (HSV) color model for detecting the human skin color, and then they detected the hand shape by applying edge detection. However some images were not detected correctly because of light conditions, irregular background and geometric variations. Wang et al. [5] proposed a Hand sign detection system was proposed based on compact hand extraction. They advanced a method to eliminate dark background. A YCrCr model was used for extracting of the hand, and then they line up and recombine fingers. Finally, a singular value decomposition was used to decrease the effect of the illumination. An Indonesian Sign Language has been developed in real time [6], YCrCb color space was used to remove the background and detect the hand region and the skin color. To extract the hand’s features, Convexity detection algorithms were used. The classification was performed using back propagation neural network. An accuracy rate of 91.66% was achieved.
To communicate with dumb and deaf, hand gesture recognition system was proposed with android device [7]. An IPwebcom application was used to acquire hand gesture using an android device. Also a Sobel edge detector was applied in addition to morphological operation to remove the selected foreground pixels from binary image. A smart camera was introduced in [8] to recognize the gesture of the hand. There are different modules in the proposed system: the first module defines the hand as it is using a mask color algorithm and also extracts it from the background using the static image way. After that, it was to check hand gesture. Actually, the number of finger was determined by finding the contour of the hand. Using the number of fingers and angle difference between fingers, the fingertips and palm center of the hand could be extracted. The success in this system is high, but the response time didn’t face the real time constraint. To solve this problem, mobile cloud computing was used to reduce processing time. A face detection and hand gesture recognition based on HCI system was presented in [9]. They recognized the position of mouth and the eyes, and estimate the pose of the head, facial center was used. The work presented two approaches: automatic hand gesture segmentation area and normalization of orientation. The accuracy rate of this method was 93.6%. In 2014, Farook and Ali [10] proposed three techniques for hand features extraction called curvature of perimeter, k-curvature, and convex hull. Also a real time hand gesture recognition system for American Sign Language (ASL) was proposed in [11]. The system acquired hand gesture images of American Sign Language by a black background that is taken with video camera of a mobile phone. Five features extraction algorithms was used such as fingertips [12, 13], eccentricity [14], elongation, automatic pixel segmentation [9, 15], and rotation. An algorithm was proposed which combined K-curvature and convex hull to detect fingertips correctly. For recognition, feed forward back propagation Neural Network was used to classify 37 hand sign gestures. The accuracy rate of gesture recognition was 94.32% in real time.

A useful solution was introduced by using kinect camera in static hand gestures recognition to extract data from depth image [16]. Depth–based data are very robust to solve problems like position and orientation of the hand, also illumination conditions than color–based data. In addition they can capture the best poses of the hand and fingers, so that the process of the recognition is easier. A simple approach was used to segment the hand region. Then, applied filtering and edge detection procedures. After that, moment invariant and DCT were used to extract features. Finally, ANN was trained to classify sign language. The result explained that using the fused features was more effective than the stand – alone approaches. In [17], a real time hand gesture based on depth data collected using kinect sensor is described. The system used k-curvature algorithm to find fingertips over the contour, and Dynamic Time Wrapping (DTW) was used to select a candidate gesture as well as to distinguish gestures by comparing them with a sequence of prerecorded reference gestures. The system achieved an average efficiency of 92.4%. Framework recognition of sign languages has been proposed based on kinect in [18]. The framework contains the extraction of features, modeling and recognition. In the feature extraction stage, they offered an improved form context feature that captured well the spatial and temporal information regarding the functionality of the appearance. Histogram of Oriented Gradients (HOG) [19] with Principal Component Analysis (PCA) function was used in the modeling phase. They proposed a method for adaptive states inspired by the change in the forms of the hands, rather than using fixed hidden states in the HMM, which achieved better results. Microsoft’s Kinect camera used to capture image called depth map or depth image which can simply segment and identify the acquired image, and track the image in 3D space. However, the problem is that it is extremely expensive.

Hand gesture recognition system running on mobile device was proposed in [20]. They built client server system applied to ASL alphabet recognition. The system using the HOG algorithm was used to detect appropriate image by way of comparing the input image acquired by the camera. The result of the experiment proved the strength of this system compared to the variation in rotation.

Another method of static gesture recognition Using two-dimensional Zernike moments (2DZM) was presented in [21], which are considered effective features when pattern in the images have distortions due to rotation, scaling or the angle of vision. The nearest classifier was used on discriminating ZM (DZM) to recognize the posture of the hand in an efficient manner in terms of calculations. The experimental results shown that DZM method improved the recognition accuracy compared to the existing methods. Moreover, Huong et al. [22] proposed a hand gesture recognition system based on Vietnamese Sign Language (VSL) in uniform background. They used PCA to recognize hand gesture. The results showed that Euclidean distance achieved 90.4% and Mahalanobis distance achieved 91.5%. Other approaches for feature detection in [23] were reported. Gray Level Co–Occurrence Matrix (GLCM) [24] and Local Binary Pattern (LBP) [25] which were used for extracting texture feature. They used k-Nearest Neighbor (KNN) function for classification. The accuracy of LBP was more than GLCM, LBP achieved 94.00% and GLCM achieved 88.00%.

Still much work is needed for increasing the recognition rate. This is the aim of the present work. This paper presents a hand gesture recognition system based on various features extraction methods for American Sign Language (ASL) and evaluates the efficiency using Multiclass Support Vector Machine (SVM) and Deep Neural Network (DNN) in classification.
III. A FOUR STAGE RECOGNITION SYSTEM

The proposed hand gesture recognition system consists of four stages: image acquisition, preprocessing, feature extraction and recognition. Fig. 1 shows the block diagram of the proposed system.

Figure 1. Proposed framework of sign language recognition.

A. Database Used

The data used was obtained from the American Sign Language (ASL) [11]. Samples were collected from more than one person. There are 1850 images from 37 signs in the database, fifty images for every sign, forty images used for training phase and the rest for testing. All the ASL alphabets and numbers are involved in the signs, as shown in Fig. 2.

B. Preprocessing

After collecting the required images, the images were preprocessed to enhance the images and to suppress any background noise, if present. The image was captured with solid background (any dark background).

- First, resize images to 260 × 260 pixels, and then convert the RGB image (color image) to a gray scale image as shown in Fig. 3.
- The image was then transformed to binary by the threshold of the Global histogram using the Otsu method as shown in Fig. 3.
- The median filter is applied to eliminate noise and preserve the edges as shown in Fig. 3.
- Fill holes and smooth the edges, using morphological bridge and diag. Then cut out the rest of the part of the image as shown in Fig. 3.

Figure 3. Steps of preprocessing to detect hand region.
The hand’s fingers need to be aligned vertically upwards. As a result, the rotation of image will round from 0 to 360 degrees based on the position of hand’s wrist. It was searched for 15 white pixels following as a minimum at the end of the image in order to put the wrist. In the case of searching for the white pixels at the end of the image, so the rotation will not be needed. In the opposite case, the 90 degrees of the image will be checked and rotated on the clock over again. By this way, the cycle will continue till finding the hand wrist. As shown in Fig. 3. This process is done if rotation used as feature.

After using the above steps to determine the hand region, the binary image was used as a mask to extract the hand from the original image. Then it was converted to gray image, shown in Fig. 4.

C. Feature Extraction

The next step to extract features that can identify each image. Distinct features algorithms such as fingertip finder, automatic pixel segmentation, eccentricity, elongation, rotation, Histogram of Oriented Gradients, and Local Binary Pattern which are used for feature extraction.

1) Fingertip finder algorithm: This step was implemented by combining K curvature algorithm with convex hull algorithm. The convex hull is used to compose a polygon around the extracted hand region, also The K curvature algorithm is used to find an angle between two points [11]. A total of 11 features were extracted from each hand.

2) Automatic pixel segmentation: According to the algorithm steps [9, 15], the image is divided to 16 blocks. There are 256 pixels in each block. Then white pixels have been calculated in each block. Accordingly, there are 16 feature vectors was produced from the number of white pixels.

3) Eccentricity: The eccentricity of the shape is the ratio of the major axis to the minor axis, see [11].

4) Elongation: Elongation of a region is defined as the ratio between the length and the width of the minimum bound rectangle of the region [11].

5) Rotation: In ASL, there are some signs that are very similar. The only difference between them is the angle of rotation [11].

6) Local binary pattern (LBP) algorithm: LBP is easy algorithm, but very efficient texture operator that labels the pixels of an image by the threshold of the neighborhood of every pixel and considers the result as a binary number. Due to its discriminative power and computational simplicity, the texture operator LBP has become a popular technique in diverse applications. It can be seen as a unifying approach to the traditionally divergent statistical and structural models of texture analysis. Perhaps the most important advantage of the LBP function in world of nature applications is its strength against the monotonous changes in gray scale caused, for example, by variations in lighting. Another important property is its computational simplicity, which makes it possible to analyze images in challenging environments in real time.

The LBP method has proved to outperform many existing methods, including the linear discriminant analysis and the principal component analysis. In order to deal with textures at different scales, the LBP operator was later extended to use neighborhoods of different sizes. As a set of sampling points evenly spaced on a circle centered at the pixel to be labeled by defining the local neighborhood allows any radius and number of sampling points. When a sampling point does not fall in the center of a pixel, bilinear interpolation was employed in the LBP method where each pixel is replaced by a binary pattern that is derived from the pixel’s neighborhood.

Instead of converting the original image to gray and apply LBP. First, the hand region was determined. Then the hand was extracted from original image. The image was then converted into gray, and the LBP algorithm was applied as shown in Fig. 5.
7) Histogram of Oriented Gradients (HOG): The basic idea of the algorithm is that local shape information often well described by the distribution of intensity gradients or edge directions even without precise information about the location of the edges themselves. In practice this is implemented by dividing the image into small sub-images called cells, and for the pixels within each cell accumulating a histogram of edge orientation. The cell can be rectangular or circular. The combined histogram entries are used as the feature vector describing the object. The advantages of this algorithm are capturing edge or gradient structure that is very characteristic of local shape, also relatively invariant to local geometric and photometric transformations.

After hand region is detected by binary image, hand can be extracted from original image then HOG Algorithm can be applied as shown in Fig. 6.

D. Classification and Fusion

Having extracted the main image features, the next step is classification where selected features are utilized as inputs to the designed classifier. Two types of classifiers were designed: Multiclass support vector machine (MSVM) and Deep Neural Network (DNN) as follows:

8) Multiclass support vector machine (MSVM) [26]: An SVM is a binary classifier, that is, class labels can only take two values: ±1. Multiclass SVM is useful for the classification of multiple gestures where the number of test gestures is more than two classes. We use the technique of the one against all method in which we take the training samples with the same label as one class and the others as the other class. Radial Basis Function Kernel (RBF) kernel was implemented because it provides better results. This kernel linearly maps samples in a higher dimensional space so that, unlike the linear kernel, it can handle the case when the relationship between class labels and attributes is not linear [27]. To estimate the generalization error, cross validation with 10 fold was used as shown in table 1.

9) Deep learning Neural Network (DNN): An autoencoder technique (AE) was used to identify American numbers and alphabets. An autoencoder (AE) is an unsupervised learning algorithm which is being used to efficiently encode the data set in order to reduce dimensionality [28]. Recently, EAs have been used to learn generative data models. The input data is first converted into an abstract representation that encoder function converted back to the original format. To be more specific, it is trained to encode the input in some representation in order that the input can be reconstructed from that representation. Stacked autoencoder was used as supervised learning algorithm by using two autoencoder networks. At the first one, hidden layer size of neurons for all features vector as shown in table II, and decoder transfer function is log sig. At the second autoencoder network, it was trained using the feature from the first autoencoder and the hidden layer size of the neurons shown in table II.

In both tables I and II below:
* Refers to Fingertip Finder, Automatic Pixel Segmentation, Eccentricity, Elongation, and Rotation,
A Refer to Local Binary Pattern (LBP), # Refers to Histogram of Oriented Gradients (HOG).

TABLE I. THE GENERALIZATION ERROR OF SVM NETWORK TRAINED WITH FEATURES

<table>
<thead>
<tr>
<th>Feature</th>
<th>*</th>
<th>^</th>
<th>#</th>
<th>* + ^</th>
<th>* + #</th>
<th>* ^</th>
<th>+ #</th>
<th>* ^ + ^ #</th>
</tr>
</thead>
<tbody>
<tr>
<td>Error</td>
<td>0.0122</td>
<td>0.0331</td>
<td>0.0162</td>
<td>0.0128</td>
<td>0.0122</td>
<td>0.0122</td>
<td>0.0122</td>
<td></td>
</tr>
</tbody>
</table>

TABLE II. THE HIDDEN LAYER SIZE OF THE NEURONS FOR TWO AUTOENCODER NETWORKS

<table>
<thead>
<tr>
<th>Feature</th>
<th>*</th>
<th>^</th>
<th>#</th>
<th>* + ^</th>
<th>* + #</th>
<th>* ^</th>
<th>+ #</th>
<th>* ^ + ^ #</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st Weight</td>
<td>20</td>
<td>150</td>
<td>150</td>
<td>150</td>
<td>1000</td>
<td>150</td>
<td>1000</td>
<td></td>
</tr>
<tr>
<td>2nd Weight</td>
<td>30</td>
<td>100</td>
<td>100</td>
<td>400</td>
<td>100</td>
<td>400</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

E. The Feature Vectors

As shown in Fig. 7. Each feature vector was trained separately then they were combined and trained together finally score level fusion was carried out. After extracted fingertip finder, automatic pixel segmentation, eccentricity, elongation, and rotation. The vector feature contains 30 features. For LBP, the feature vector containing 2124 features. Using HOG, the feature vector contains 2000 features. All feature vectors were recognized using SVM and autoencoder.

10) Feature level fusion: Feature level fusion can significantly enhance the performance of hand recognition system. Feature level fusion can be found by concatenating the feature sets taken from hand gestures images [29]. By concatenating two feature vectors, a new feature vector has been created. The target is to use this new feature set for classification. As shown in Fig. 7.

11) Score level fusion: Score level fusion [29] is an effective method to integrate various single classifiers when a set of classifiers output are created. Score level fusion is calculated using maximum technique [30] to achieve the final recognition results, as shown in Fig. 7.

IV. RESULTS AND DISCUSSION

The proposed method was implemented in MATLAB 2016 b. A dataset is used to test the system from [11]. Ten hand images were used for each sign. As previous steps, preprocessing and feature are extracted from the images, and then tested in previous learned network. Fig.8 and 9 show the experimental results to recognize 37 sign of ASL hand gesture.
V. CONCLUSION

In our research reported here we carried out Automatic hand sign language recognition based on computer vision to distinguish 37 signs of American numbers and alphabets. Different algorithms were applied to extract the features from visible gestures. These included: Local Binary Pattern (LBP), Histogram of Oriented Gradients (HOG), fingertip finder, pixel segmentation, eccentricity, elongation, and rotation. In addition, feature fusion and score fusion were applied to achieve high accuracy rate in the recognition system. By a mobile video camera interconnected to a computer, some image frames were taken, and all were tested by our trained network multiclass Support Vector Machine and Autoencoder algorithm. According to the results, we can train the system to define more American static sign language for the signs of hand while still maintaining high accuracy. The results demonstrated that we can build on the framework to acknowledge the dynamic sign language of the hand as well as identify the hand for words.
ACKNOWLEDGEMENTS

The authors would like to thank Md. Mohimun Islam, Shahjalal University of Science and Technology, for providing us with the database used in the present work.

REFERENCES


AUTHORS’ PROFILES

Aida Abd El-Mohsen obtained the BS degree from the Department of Information Technology, Computers and Information Science (FCIS), Mansoura University, Egypt, in 2010. From 2011 to 2012, she worked as Research assistant in FCIS.
Eman M. El- Daydamony received her BS, MS, and PhD degrees in electrical communications from the Faculty of Engineering, Mansoura University, Egypt, in 1998, 2003, and 2008, respectively. From 1998 to 2002, she worked as a teaching assistant. From 2003 to 2007, she worked as assistant teacher, and from 2008 to 2010, she worked as a lecturer. From 2010 to 2017, she worked as a lecturer in the Faculty of Computers and Information Sciences (FCIS), Mansoura University, Egypt. Since 2018, she has worked as an associate professor at FCIS.

Fatma E. Z. Abou-Chadi: obtained her PhD from Imperial College of Science and Technology, UK. She is an IEEE Senior Member since 1995, and a Member of the Society of IEEE Women in Engineering since 2000. She was the recipient of the IEEE Region 8 Clementina Saduwa Award for Year 2013. She was the Dean of the Faculty of Computers & Information Sciences, Mansoura University during the period 2007-2012. She is now the Chairwoman of the Department of Electrical Engineering, Faculty of Engineering - The British University in Egypt.