Annals of Emerging Technologies in Computing (AETiC)
Vol. 4, No. 4, 2020

Hand Gesture-based Sign Alphabet Recognition and Sentence Interpretation using a Convolutional Neural Network

Md. Abdur Rahim[1], Jungpil Shin[1,*], Keun Soo Yun[2,*]

[1]School of Computer Science and Engineering, The University of Aizu, Japan
rahim_bds@yahoo.com; jpshin@u-aizu.ac.jp
[2]Department of Computer and Information, Ulsan College, Ulsan, Korea
ksyun@uc.ac.kr
*Correspondence: ksyun@uc.ac.kr; jpshin@u-aizu.ac.jp

Received: 3rd June 2020; Accepted: 25th August 2020; Published: 1st October 2020

Abstract: Sign language (SL) recognition is intended to connect deaf people with the general population via a variety of perspectives, experiences, and skills that serve as a basis for the development of human-computer interaction. Hand gesture-based SL recognition encompasses a wide range of human capabilities and perspectives. The efficiency of hand gesture performance is still challenging due to the complexity of varying levels of illumination, diversity, multiple aspects, self-identifying parts, different shapes, sizes, and complex backgrounds. In this context, we present an American Sign Language alphabet recognition system that translates sign gestures into text and creates a meaningful sentence from continuously performed gestures. We propose a segmentation technique for hand gestures and present a convolutional neural network (CNN) based on the fusion of features. The input image is captured directly from a video via a low-cost device such as a webcam and is pre-processed by a filtering and segmentation technique, for example the Otsu method. Following this, a CNN is used to extract the features, which are then fused in a fully connected layer. To classify and recognize the sign gestures, a well-known classifier such as Softmax is used. A dataset is proposed for this work that contains only static images of hand gestures, which were collected in a laboratory environment. An analysis of the results shows that our proposed system achieves better recognition accuracy than other state-of-the-art systems.

Keywords: ***Convolutional neural network; Human-computer interaction; Hand gesture; Otsu method; Sign language***

---

1. Introduction

Sign language (SL) involves movements of different parts of the body, for example the face and hands, which deaf and hearing-impaired people use to interact with hearing people. Hand gestures have particular importance in the recognition of SL, which has its own structure and grammar, and changes with the fluency of signing, involving the use of different types of movements such as static and dynamic. In this work, we use only static images for all American Sign Language (ASL) gestures, and present an ASL alphabet recognition system that recognizes the different hand gestures, assesses various approaches to the recognition of the signs and eventually interprets them into a meaningful sentence. We develop a sign dataset based solely on the shape and orientation of the hand, and movements of the face and head are not considered. Many researchers have contributed to the study of human-computer interaction with the aim of improving communication between deaf people and the general public. McKee *et al*. [1] developed a health literacy system for

Md. Abdur Rahim, Jungpil Shin and Keun Soo Yun, "Hand Gesture-based Sign Alphabet Recognition and Sentence Interpretation using a Convolutional Neural Network", *Annals of Emerging Technologies in Computing (AETiC)*, Print ISSN: 2516-0281, Online ISSN: 2516-029X, pp. 20-27, Vol. 4, No. 4, 1st October 2020, Published by International Association of Educators and Researchers (IAER), DOI: 10.33166/AETiC.2020.04.003, Available: http://aetic.theiaer.org/archive/v4/v4n4/p3.html.

ASL users to establish adequate health literacy connections between deaf ASL users and English speakers [1]. In [2], the authors described the importance of SL recognition with respect to hand and face signs. However, the recognition of sign gestures with varying levels of illumination and complex backgrounds remains a major concern. An ASL character recognition system was proposed in [3], although the authors considered only five ASL characters. Hand gesture-based recognition systems for isolated sign words were proposed in [4], in which feature fusion techniques were used to detect the sign words. In [5], the authors discussed the issue of SL recognition based on hand size and hand movement information via wearable devices. Glove-based gesture recognition requires the user to wear a data glove in order to generate gesture-related information, and this can be uncomfortable and unhygienic. In [6], a hidden Markov and depth sensor device-based classification system was proposed to learn the sign gestures. However, there is no clear explanation for fingertip detection in different aspects of light illumination. The fingerspelling alphabet of the ASL recognition system was presented in [7]. Here an SVM technique was used to classify the sign. However, it is unable to detect when the fingers overlap. In [8], the authors proposed preprocessing the HSV color space and applied hand gesture segmentation based on skin pixels. However, the proposed system is unable to reduce noise from the input image. Depth sensor-based ASL recognition was proposed [9]. This system was developed using 26 characters and 10 numbers. However, leap motion has a large number of features, therefore, it is difficult to identify the effective features.

In the present paper, a non-wearable device is used to collect input from 26 ASL alphabet gestures and to pre-process it using the proposed method. Feature fusion with a CNN is used to extract the features. Our system takes various features of the gesture image as input and executes a convolution process. The features are then analyzed for gesture classification.

The rest of this paper is structured as follows: Section 2 briefly discusses the details of the image dataset, the pre-processing of input images, feature extraction, and the classification processes of the proposed system. Section 3 explains the experimental results and their implications. Section 4 summarizes this work and gives an outline of future work.

2. Proposed System

In this study, the proposed system is used to detect hand gestures and to interpret a meaningful word or sentence from a continuous performance of ASL gestures. The system consists of four main steps: data collection, segmentation, feature extraction, and finally classification. Fig. 1 illustrates the proposed ASL alphabet recognition and sentence interpretation system.

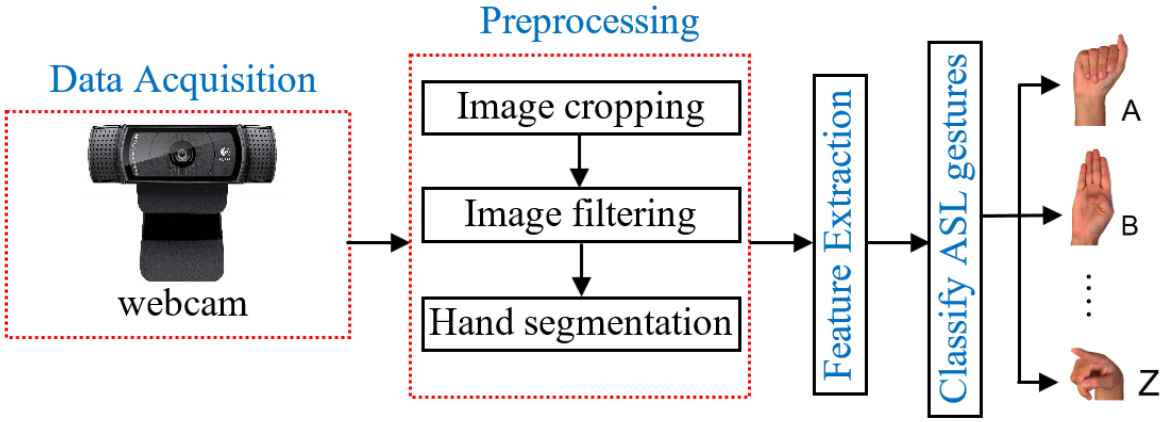

Figure 1. Basic architecture of ASL alphabet recognition and sentence interpretation

2.1. Description of the Image Dataset

In order to create an image dataset, data were taken from a live camera, and hand gesture images were obtained from the region of interest (ROI) and used as input. In this approach, no accessories such as gloves or wearable devices are needed. The dataset contained 26 ASL gestures representing the alphabet, and the images had a resolution of 50 x 50 pixels. A total of 23,400 images

were collected for the 26 ASL gestures, with 900 images captured for each gesture. In this work, each image contained hands at different locations, and these images were obtained under different lighting conditions and against differing background environments. Fig. 2 shows examples of the images in the ASL dataset.

Figure 2. Example of ASL dataset images from A to Z.

2.2. Image Pre-processing

In this system, an image of a static hand gesture was used as input, which was captured via a webcam. We captured these images from the region of interest (ROI) locations. The specific area of the frame expressed as the ROI is considered to identify hand gestures by which unnecessary areas of the video frame can be ignored. A Gaussian blurring technique was used to make the image smoother and to reduce the noise in the input. We then applied the Otsu method to pre-process the input image; this provides an intuitive way of automatically selecting a threshold based on the statistics of the image's histogram, which acts as a form of image segmentation. However, the quality of thresholding is dependent on the selection of the threshold intensity, as the optimal threshold intensity is determined based on a histogram of the image. The proposed method reduces the intra-class intensity [10] and determines the marginal threshold using Equation (1).

$$\sigma_w^2(t) = q_1(t)\sigma_1^2(t) + q_2(t)\sigma_2^2(t) \quad (1)$$

where the class probabilities are q_1 and q_2, the threshold is t, and the weighted sum of the variance is $\sigma_w^2(t)$. The probabilities of the class were calculated from the histogram. The smallest results of $\sigma_w^2(t)$ is applied to the threshold. To calculate this value, we created a normalized histogram that defines *P(i)*. This *P(i)* gives the percentage of intensity of the pixels *i*. In addition, we calculated weights, mean, and variances for each class. Table 1 represents the descriptions of calculation process of weights, means, and variances.

Table 1. Calculation process of weights, means, and variances.

Function	Equations	Description
Weights	$q_1(t) = \sum_{i=0}^{t} P(i)$, $q_2(t) = \sum_{t=t+1}^{L-1} P(i)$	Left class intensity values *0* to *t*, right class intensity values *t+1* to *L-1*.
Means	$\mu_1(t) = \sum_{i=0}^{t} \frac{i*P(i)}{q_1(t)}$, $\mu_2(t) = \sum_{t=t+1}^{L-1} \frac{i*P(i)}{q_2(t)}$	Probability and weight are used to measure mean.
Variances	$\sigma_1^2(t) = \sum_{i=1}^{t}[i-\mu_1(t)]^2 \frac{P(i)}{q_1(t)}$, $\sigma_2^2(t) = \sum_{t=t+1}^{L-1}[i-\mu_2(t)]^2 \frac{P(i)}{q_2(t)}$	Weights and means are used to measure variances

2.3. Feature Extraction and Classification

The use of deep networks is remarkable in technological advances that can be used to classify images and to identify and distinguish different aspects of images [11]. It consists of one or more convoluted layers that are primarily used for image processing, classification, and other co-related data processing. In this perspective, feature extraction is used to obtain significant information by analyzing the input image via the proposed image processing technique. The extracted feature vectors are then fed to the classification process. However, these features must contain contextual information from the input, as this is used to categorize them from other gestures to specific

gestures. We use a convolutional neural network (CNN) to extract the properties of the hand gestures, as shown in Fig. 3. The feature is extracted step by its convolutional layers. Although feature data can be fed directly to deep convolutional networks for training and testing, we used filtered and segmented images as the input in this architecture. We used a kernel function to create a feature map and padding to prevent the features from shrinking, and max-pooling was also applied to reduce the size of the features and retain important information. In the proposed architecture, fusion of features occurs at the fully connected layer, producing the output of a Softmax classification.

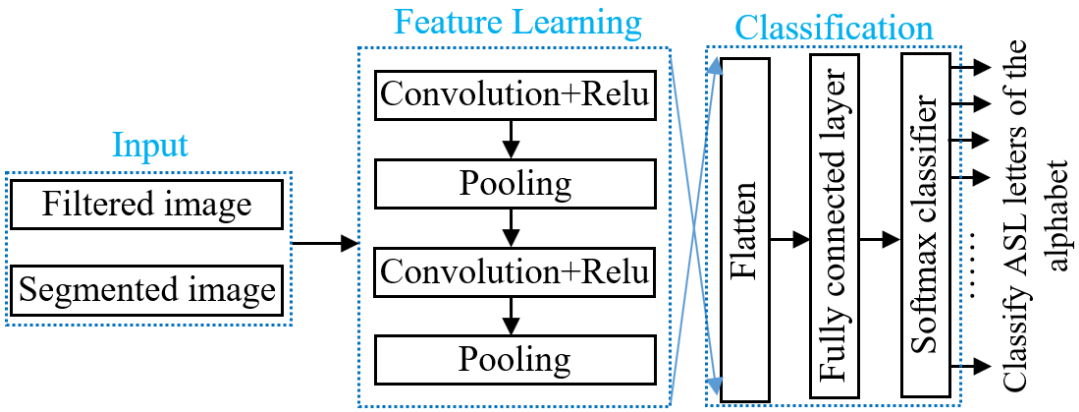

Figure 3. Proposed CNN architecture.

3. Experimental Results and Analysis

In this section, we evaluate our ASL recognition system based on our dataset[1]. Two types of tests were performed to evaluate the recognition and interpretation of sentences using the ASL alphabet. The training and test datasets contained hand gestures performed by the same individual, and different individuals performed real-time ASL alphabet recognition and sentence interpretation based on trained model.

3.1. Hand Gesture Segmentation

We segmented the hand gestures from the input image. To avoid the presence of unnecessary background and noise, the input image was filtered and the hand gesture images were segmented. These two types of images were used as input to the proposed model. Fig. 4 shows examples of segmented images from the image dataset.

Figure 4. Examples of hand gesture segmentation of input images.

3.2. ASL Recognition and Sentence Interpretation

In order to recognize the ASL alphabet and interpret meaningful sentences, we evaluated the different hand gestures. The CNN was trained on the entire dataset, and the architecture was evaluated based on two different types of input data: (i) input data containing grayscale filtered

[1] https://www.u-aizu.ac.jp/labs/is-pp/pplab/ASL_DATASET.zip

images; and (ii) input data containing images with grayscale segmentation. We used a built-in dataset containing 26 ASL gestures (letters of the alphabet) with a total of 23,400 images. We used 70% of the images for training and 30% for testing, meaning that a total of 46,800 images (i.e. both filtered and segmented images) were used, with 32,760 used for training and 14,040 for testing. The average recognition accuracy of the ASL alphabet is illustrated in Fig. 5. The confusion matrix for the accuracy of recognition of ASL gestures is presented in Fig. 6. Some misrecognition occurred using this system due to the presence of similar shapes, as illustrated in Fig. 7. The highest accuracy was seen for the letter 'Y', and the lowest for 'M'. Our system achieved an average of 96.59% recognition accuracy. Table 2 shows a comparison of the accuracy of our method with state-of-the-art alternatives. We have achieved better accuracy in our method than the accuracy mentioned in [6], [12] and [13]. In 3D images of the hand, the thumbs are often interspersed with the other finger, and the number of training information can affect the accuracy of the classification [6, 12]. The reported accuracy of [6], [12], and [13] are 86.1%, 96.15%, and 94.2%, respectively. In this paper, we proposed segmentation techniques and introduced the fusion of features that enhance the validity of recognition.

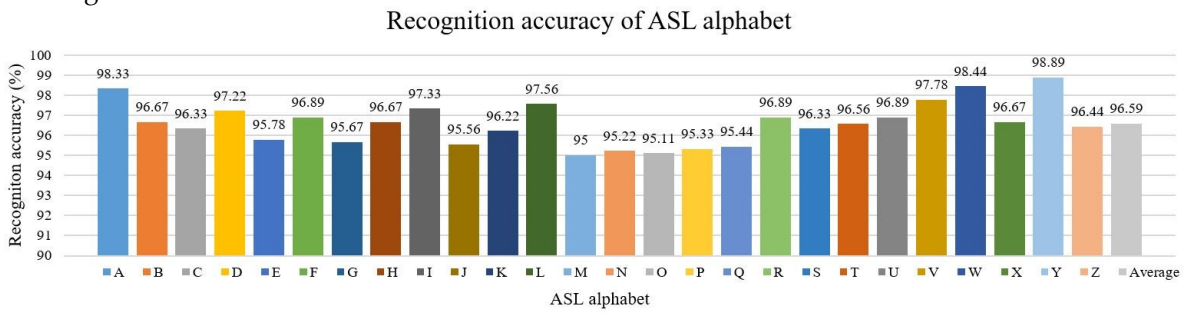

Figure 5. Recognition accuracy of the ASL alphabet

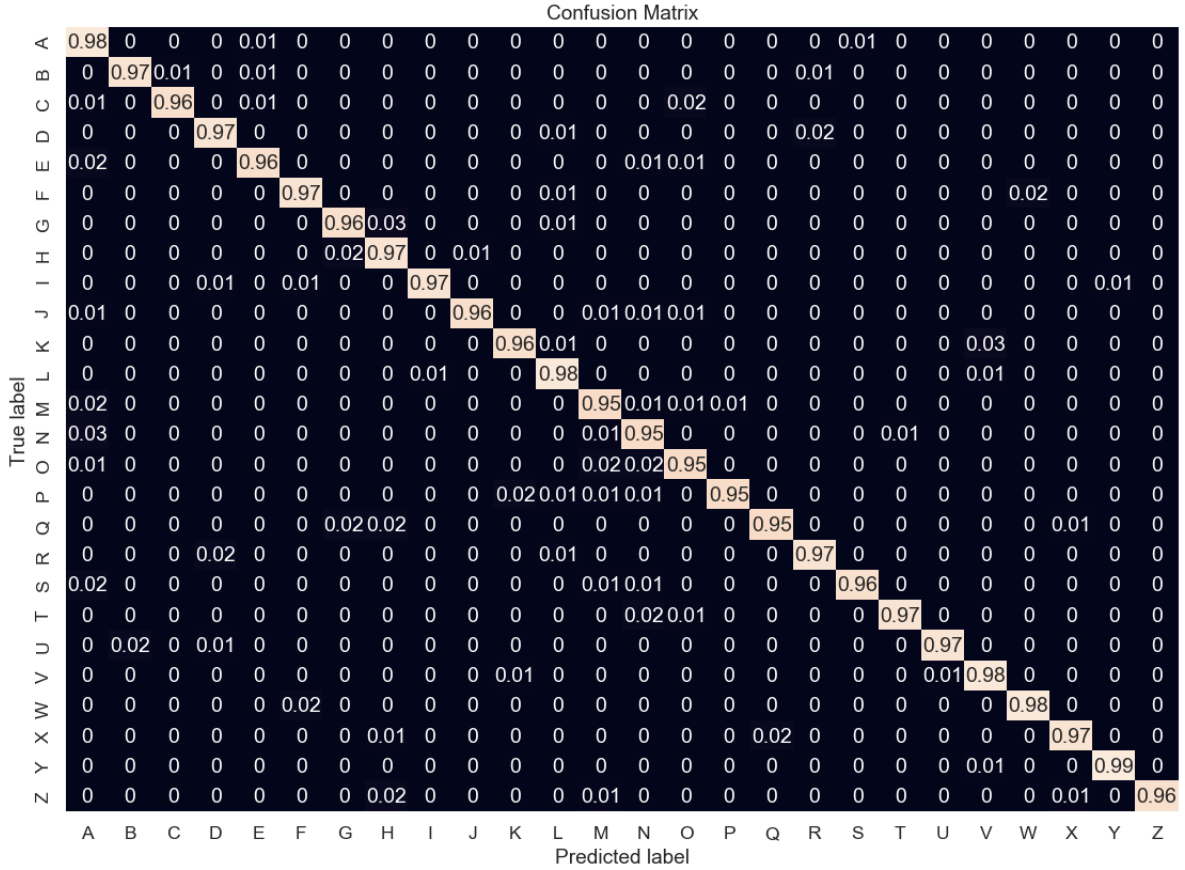

Figure 6. Confusion matrix for the sign alphabet

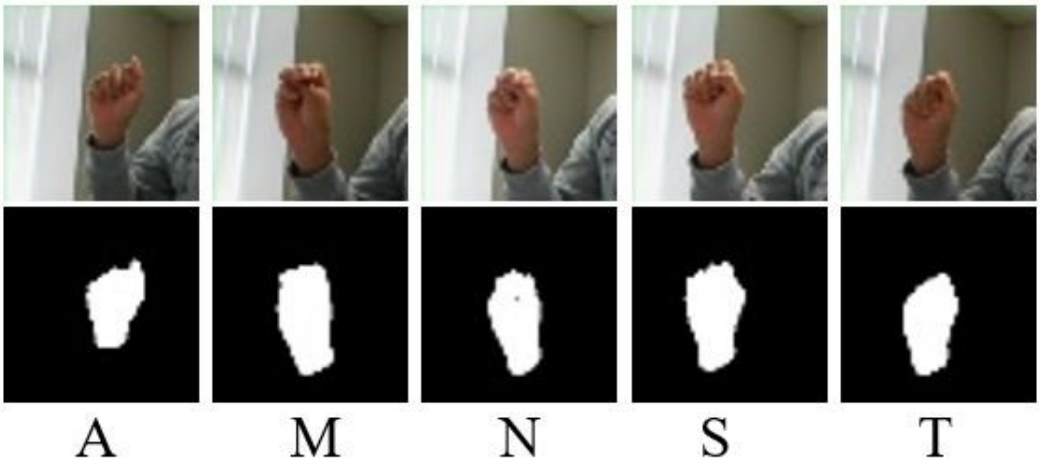

Figure 7. Examples of similar shapes in the ASL alphabet

Table 2. Dataset descriptions and comparison of recognition accuracy with other methods

References	Subjects	Classification Method	Number of gestures	Total number of gestures	Recognition Accuracy (%)
Ref. [6]	12	Hidden Markov classification	24 ASL	2880 (each gesture 10 times)	86.1
Ref. [12]	40	HOG Feature and SVM	10 ASL	650 images for 10 classes (selected from database)	96.15
Ref. [13]	9	CNN	25 ASL	675 samples (each gesture 3 times)	94.2
Proposed method	3	CNN and Softmax	26 ASL	23,400 (each gesture 300 times)	96.59

We also tested this system with different individuals in real time, in terms of recognizing the ASL alphabet and interpreting these continuous sign gestures to give meaningful sentences. Six respondents were asked to perform sign gestures at the ROI location. The user continued to perform gestures for creating sentences like 'MY UNCLE DIVORCE', 'SHE HAPPY BABY'. The system evaluated each gesture after pre-processing, feature extraction, and classification. We classified the performed gestures based on the trained model. Fig. 8 shows the average recognition accuracy for each user, and Fig. 9 presents an example of a sign gesture being performed continuously in real time.

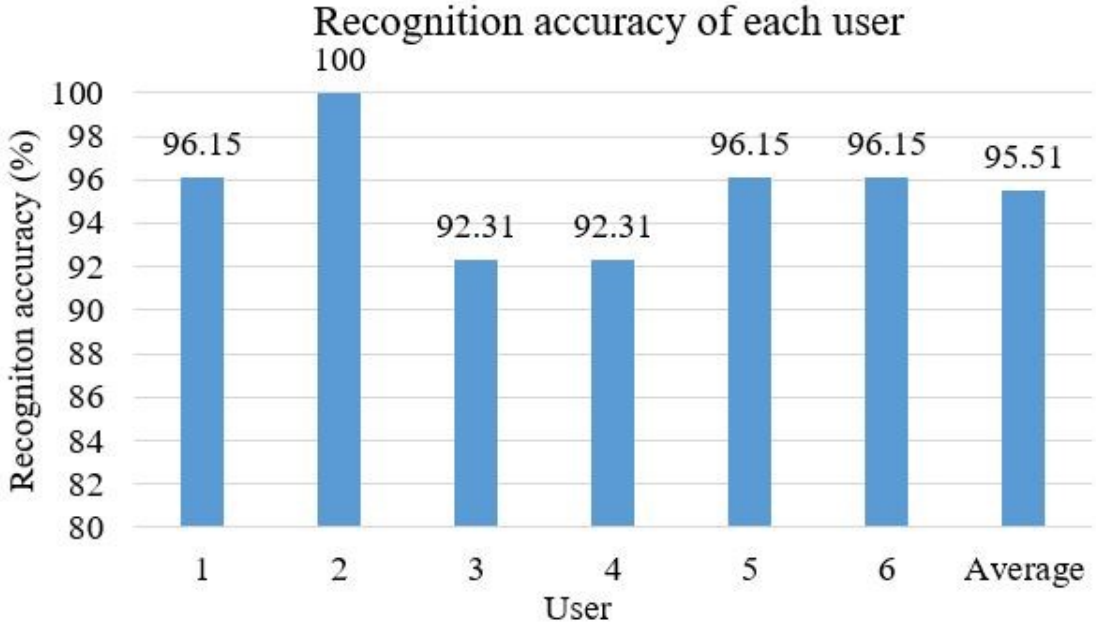

Figure 8. Average recognition accuracy for each of six users

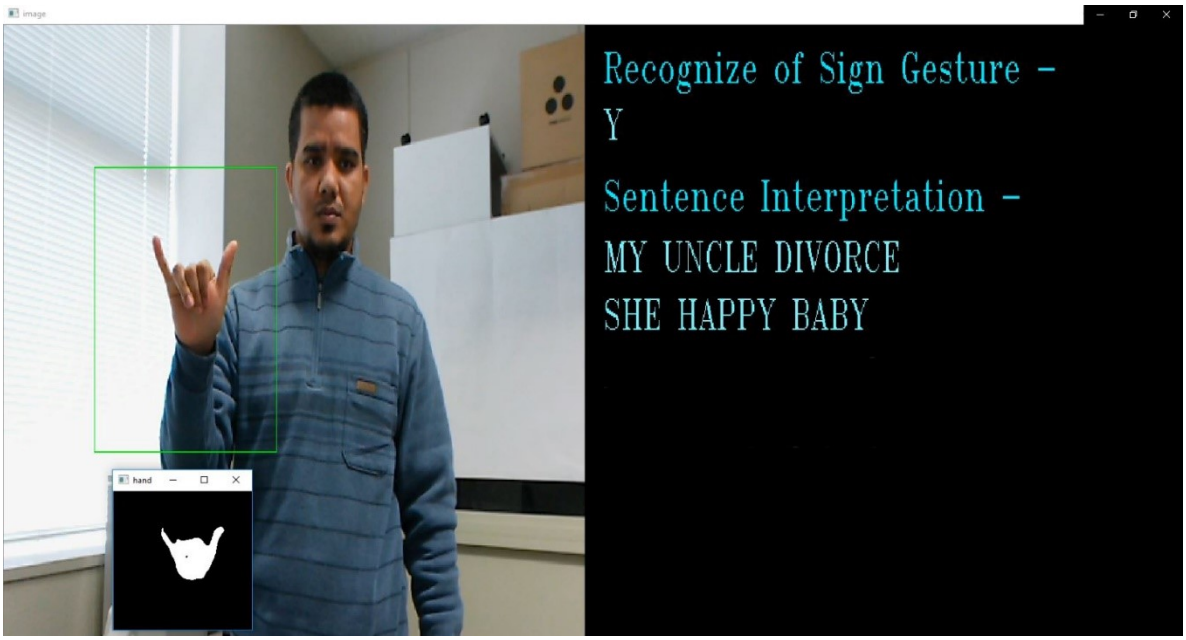

Figure 9. Example of ASL alphabet recognition and sentence interpretation

4. Conclusion

In this paper, we introduce a stable hand gesture system for ASL recognition that is able to detect hand positions, translate gestures into text and interpret meaningful sentences from continuously performed gestures. A dataset was created using a low-cost webcam and pre-processed images. A Gaussian blurring technique was used to filter the images in the dataset and the Otsu method was applied to determine the standard threshold value for segmentation of the hand. The filtered and segmented images were then fed into the feature extraction process. A two-channel CNN architecture was proposed to extract the features from the input images, and fusion of the features was implemented in the fully connected layer. A Softmax classifier was applied to classify the gestures. The experimental results show that the recognition accuracy for the ASL alphabet was 96.59%, i.e. better than other state-of-the-art systems. This system can also create meaningful sentences from the continuous performance of gestures. To improve our scheme in future work, we intend to collect more data on dynamic hand gestures and to enrich it with more signs/words.

References

[1] McKee, M.M.; Paasche-Orlow, M.K.; Winters, P.C.; Fiscella, K.; Zazove, P.; Sen, A.; Pearson, T. Assessing health literacy in deaf American Sign Language users. Journal of Health Communication 2015, Vol. 20, pp. 92-100, DOI: 10.1080/10810730.2015.1066468.

[2] Koller, O.; Forster, J.; Ney, H. Continuous sign language recognition: Towards large vocabulary statistical recognition systems handling multiple signers. Computer Vision and Image Understanding 2015, Vol. 141, pp.108-125, DOI: 10.1016/j.cviu.2015.09.013.

[3] Jiang, X.; Ahmad, W. Hand gesture detection based real-time American Sign Language letters recognition using support vector machine. In Proceedings of the 2019 IEEE International Conference on Dependable, Autonomic and Secure Computing, International Conference on Pervasive Intelligence and Computing, International Conference on Cloud and Big Data Computing, International Conference on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pp. 380-385, DOI: 10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00078.

[4] Rahim, M.A.; Islam, M.R.; Shin, J. Non-touch sign word recognition based on dynamic hand gesture using hybrid segmentation and CNN feature fusion. Applied Sciences 2015, Vol. 9, No. 18, 3790, Available: https://doi.org/10.3390/app9183790.

[5] Ahmed, M.A.; Zaidan, B.B.; Zaidan, A.A.; Salih, M.M.; Lakulu, M.M.B. A review on systems-based sensory gloves for sign language recognition state of the art between 2007 and 2017. Sensors 2018, Vol. 18, No. 7, p. 2208, Available: https://doi.org/10.3390/s18072208.

[6] Vaitkevičius, A.; Taroza, M.; Blažauskas, T.; Damaševičius, R.; Maskeliūnas, R.; and Woźniak, M. Recognition of American sign language gestures in a virtual reality using leap motion. Applied Sciences 2019, Vol. 9, No. 3, pp. 445, Available: https://doi.org/10.3390/app9030445.

[7] Quesada, L.; López, G.; Guerrero, L. Automatic recognition of the American sign language fingerspelling alphabet to assist people living with speech or hearing impairments. Journal of Ambient Intelligence and Humanized Computing 2017, Vol. 8, No. 4, pp.625-635, Available: https://doi.org/10.1007/s12652-017-0475-7.

[8] Raheja, J.L.; Mishra, A.; Chaudhary, A. Indian sign language recognition using SVM. Pattern Recognition and Image Analysis, 2016, 26(2), pp. 434-441, Available: https://doi.org/10.1134/S1054661816020164.

[9] Chong, T.W.; Lee, B.G. American sign language recognition using leap motion controller with machine learning approach. Sensors, 2018, Vol. 18, No. 10, pp. 3554, Available: https://doi.org/10.3390/s18103554.

[10] Merzban, M.H.; Elbayoumi, M. Efficient solution of Otsu multilevel image thresholding: A comparative study. Expert Systems with Applications, 2019, Vol. 116, pp.299-309, Available: https://doi.org/10.1016/j.eswa.2018.09.008.

[11] Wu, X.Y. A hand gesture recognition algorithm based on DC-CNN. Multimedia Tools and Applications 2020, Vol. 79, pp. 9193-9205, Available: https://doi.org/10.1007/s11042-019-7193-4.

[12] Elsayed, R.A.; Abdalla, M.I.; Sayed, M.S. Hybrid method based on multi-feature descriptor for static sign language recognition. In 2017 Eighth International Conference on Intelligent Computing and Information Systems (ICICIS), IEEE, pp. 98-105, Availble: https://doi.org/10.1109/INTELCIS.2017.8260039.

[13] Huang, J.; Zhou, W.; Li, H.; Li, W. Sign language recognition using 3D convolutional neural networks. In 2015 IEEE International Conference on Multimedia and Expo (ICME), IEEE, pp. 1-6, Available: https://doi.org/10.1109/ICME.2015.7177428.