Hand and Face Segmentation with Deep Convolutional Networks using Limited Labelled Data

Ozge Mercanoglu Sincan
Computer Engineering Dept.
Ankara University
Ankara, Turkey
omercanoglu@ankara.edu.tr

Sinan Gencoglu
Computer Engineering Dept.
Ankara University
Ankara, Turkey
sinan.gencogluu@gmail.com

Mert Bacak
Computer Engineering Dept.
Ankara University
Ankara, Turkey
vlstyxz@gmail.com

Hacer Yalim Keles
Computer Engineering Dept.
Ankara University
Ankara, Turkey
hkeles@ankara.edu.tr

***Abstract*—Segmentation is a crucial step for many classification problems. There are many researchers that approach the problem using classical computer vision methods, recently deep learning approaches have been used more frequently in different domains. In this paper, we propose two segmentation networks that mark face and hands from static images for sign language recognition using only a few training data. Our networks have encoder-decoder structure that contains convolutional, max pooling and upsampling layers; the first one is a U-Net based network and the second one is a VGG-based network. We evaluate our models on two sign language datasets; the first one is our Ankara University Turkish Sign Language dataset (AU-TSL) and the second one is Montalbano Italian gesture dataset. Datasets contain background and illumination variations. Also, they are recorded with different signers. We train our models using only 400 images that we randomly selected from video frames. Our experiments show that even when we reduce the training data in half, we can still obtain satisfactory results. Proposed methods have achieved more than 98% precision using 400 frames with both datasets. Our code is available at https://github.com/au-cvml-lab/Hands-and-Face-Segmentation-With-Limited-Data.**

Keywords—gesture segmentation, face segmentation, sign language, deep learning, CNN, U-net, VGG.

I. Introduction

There has been a great deal of work in sign language recognition. Sign language consists of hand gestures, facial expressions, and body posture. In order to model a successful sign language recognition system, it is necessary to extract features robustly from hand and face regions. Therefore, some existing approaches tend to do segmentation of hands or/and face before sign recognition in order to obtain more meaningful features. Some proposals use glove-based devices. However, these approaches require the user to wear a glove with probes attached to the arm and hand of users and these often limit the movements of signers, which is not desired [1]. One of the most popular classical methods is using skin colour detection for segmentation. Drawback of this method is that skin detection is sensitive to distribution of the light in the environment. Also, skin-like objects in the background could be confused easily with face and hands [2]. In these methods, wearing long-sleeved clothing in colours that contrast with the skin colour improves the performance.

In the recent years, deep learning approaches have gained great popularity due to the performance of deep convolutional networks in different computer vision-based problems [3], [4], [5], [6]. However, deep learning approaches need large amounts of annotated training data, which makes it difficult to use them in some domains where the annotated data is limited and hard to obtain. Hand-face segmentation problem is one of them, where the image needs to be marked manually with pixel accuracy. In this work, our purpose is to train robust models using only a few training data and use the segmented regions insolving sign recognition from static images. We expect the model to segment the hand and face pixels without using any additional hardware e.g. data glove, or any constraints. We first generated a roughly marked segmentation ground truth images from sign/gesture images. We use classical image processing methods to determine candidate regions automatically and then we manually fine tune the regions. In order to segment hand and face regions for each pixel of the input image, we design two deep convolutional models that both have an encoder-decoder structure. We downsample the input images by performing a series of convolutional and maxpooling layers in the encoder part. Then, we get the desired segmented regions using upsampling in the decoder network. Our first model is based on U-Net [6] and the second one is based on VGG [7]. We evaluate our methods using two different datasets: Ankara University Turkish Sign Language (AU-TSL) dataset and Montalbano Italian Gesture dataset [8]. We train the model using images in video frames; image frames are selected randomly from the datasets.

The rest of this paper is organized as follows. We give related work in Section II. We explain the proposed method in Section III. We show the experimental results in Section IV. Finally, we conclude the paper in Section V.

II. Related Work

In sign/gesture recognition problem, segmentation can be done simultaneously with recognition or independently. In this section, we discuss the studies which perform segmentation before the classification problem.

Some existing works use wearable hardware such as data or coloured gloves. Data gloves can record hand position, velocity or angles etc. and coloured gloves simplify hand segmentation problem. In [9], data glove with tilt sensor is used to detect the bending of a finger. Furthermore, an accelerometer is used to capture the gesture motion along x, y,

978-1-7281-3789-6/19/$31.00 ©2019 IEEE

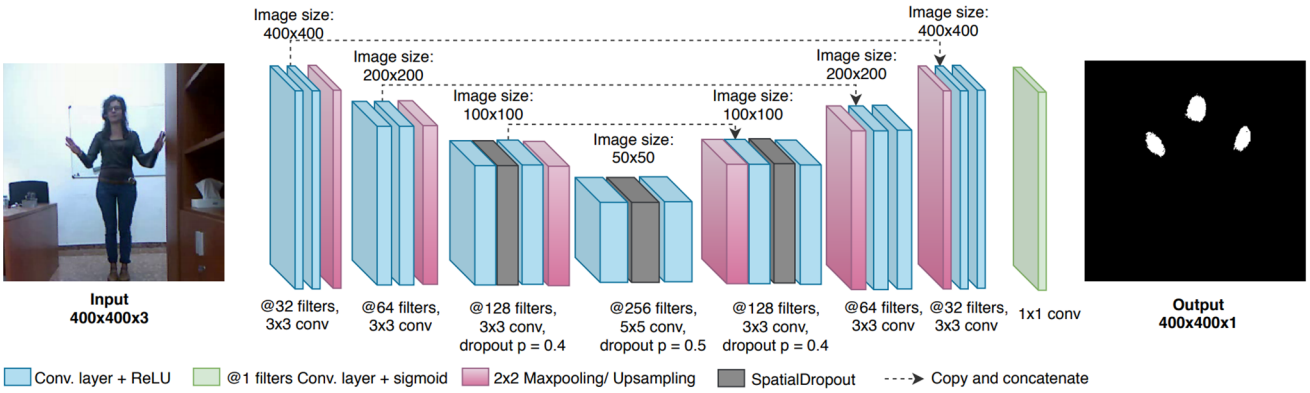

Fig. 1 Network architecture of Model1.

z axes. Furthermore, an accelerometer is used to capture the gesture motion along x, y, z axes. In [10], gesture recognition system is developed using a colour glove which consists of three different colours. Magenta is used to dye the palm, cyan and yellow are used for adjacent fingers and the rest of glove is black. HSI (hue-saturation-intensity) colour space is used for segmentation process.

Segmenting hand and face with skin colour detection is also studied in the literature. For skin colour detection, different colour spaces e.g. YCbCr, HSV (hue-saturation-value) is usually preferred instead of RGB, because they are more robust to lighting variations in the environment [11], [12]. In [12], face region is detected using Viola-Jones [13] method and the detected face is subtracted by replacing face area with a black circle. Then, skin area is detected with using HSV colour space. In order to find hands, contour comparison algorithm is used using templates of four hand gestures. Research shows that skin colour segmentation using edge detection and thresholding techniques improves the segmentation results [11].

Some early approaches try to find hand regions by using motion information. In [14], it is assumed that hand is the only moving object in the scene. In [15], sign language subunits segmentation algorithm is proposed based on hand motion speed and trajectory information. The motion speeds of the hands are calculated based on consecutive frames. Dominant hand is distinguished from their trajectory information.

Microsoft Kinect is popular in sign recognition since it provides depth information, skeleton and RGB images simultaneously. Some researchers use depth information for segmenting hand from body of the user [16], [17].

In the recent years, researchers have started to use deep learning approaches for sign language recognition. In [18], two-stage convolutional neural network (CNN) architecture is proposed for segmentation and recognition of hand gestures from static images. For segmentation stage, fully convolutional residual network is used and then Atrous Spatial Pyramid Pooling (ASPP) [19] module is used to encode multi-scale contextual information by using multiple dilation rates. OUHANDS [20] and HGR1 [21] gesture datasets are used for the evaluation of the model. In OUHANDS, the camera is hand-held and the hand is close to the camera. Some images include face, while some images do not. HGR1 dataset consists of only one hand or hand with arm. For ground truth, skin mask is provided. These datasets are not suitable for our problem since they do not include person as whole. Some researchers perform segmentation using classical methods and then use deep learning methods to classify signs. In [22], face and hands are segmented with skin colour detection method. Then, feature extraction and recognition processes are performed with deep learning techniques by using these segmented regions. In [23] and [24], sign language recognition models are proposed. Montalbano gesture dataset [8] is used for evaluation in both papers. Since Montalbano dataset are recorded with Microsoft Kinect, it also contains depth and skeleton (joint) information beside RGB. In their works, hands are cropped using joint information and cropped hands are passed to network as an extra modality. [23] is the winner of 2014 ChaLearn Looking at People Challenge [8] and [24] is the fifth team of the competition.

III. Methods

We propose two different models based on deep learning approaches. The first one is U-Net [6] based, but there are a few modifications, which we will refer to as Model1 in this paper. We have fewer blocks than U-Net and we use fewer filters. In Model1, the encoder has four blocks and the decoder has three blocks. In the encoder, blocks consist of two 3x3 convolutional layers followed by a rectified linear unit (ReLU) and 2x2 max pooling layer with a stride 2. At the first block we start with 32 feature channels and we double the number of feature channels at each block in the encoder part. In the proposed decoder, each block consists of an upsampling layer followed by concatenation with the corresponding feature map from the encoder part. The resultant features are passed from two 3x3 convolution layers with ReLU activations. Finally, model has a 1x1 convolutional layer followed by sigmoid activation function. We use spatial dropout [25] between two convolutional layers at the third, fourth blocks in the encoder and the first block of the decoder. Spatial dropout drops entire feature maps instead of individual units within feature maps. Since our model contains fully convolutional layers, and very few data spatial dropout gives better results than standard dropout [26]. The network architecture is illustrated in Fig. 1.

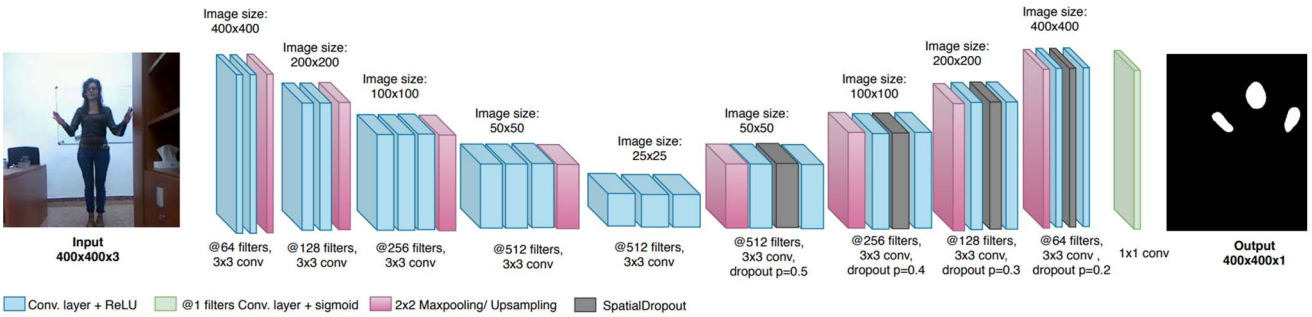

Fig. 2 Network architecture of Model2.

In our second model, namely Model2, we used VGG-16 network [7], pretrained with ImageNet [27], in the encoder part without any finetuning to our dataset. We use all the layers of VGG until the fully connected layer. In the decoder, model have four blocks that contains upsampling layer followed by two 3x3 convolutional layers with ReLU activations. Finally, model has a 1x1 convolutional layer followed by sigmoid activation function. We use spatial dropout [25] between two convolutional layers at all blocks of the decoder network with the probability values 0.5, 0.4, 0.3 and 0.2, respectively. The network architecture is illustrated in Fig. 2.

We use Adam optimization algorithm [28] with the learning rate 1e-4 for both models and batch size of 8.

IV. Experimental Results

We use two different datasets in the evaluation of our models. The first one is AU-TSL dataset that we recently have started to create, which will be complete and made public soon. We choose the words that are frequently used in the daily spoken language. We are recording our AU-TSL dataset with Microsoft Kinect V2 in three different locations. Recordings include RGB, depth, segmentation, and skeleton information. The resolution of RGB data is 1080x1920, while the others are 512x424. Our dataset contains 228 words recorded from 12 people. There are approximately 150 samples for each word, and a total of about 34,200 samples. We still continue to record data with different configurations (e.g., different background, clothing, body posture, etc.) to make the dataset more challenging. The second dataset is Montalbano Italian gesture dataset [8] that contains 20 words of Italian sign language recorded from 27 people with variations in surroundings, clothing and lighting. There are approximately 14,000 samples in total. The videos are recorded with a Microsoft Kinect. Videos have 640x480 resolution and they include RGB, depth, segmentation, and skeleton information. For both datasets, we use only RGB data. We crop the images to be square size and then resize the videos to 400x400 sizes by keeping the people on the horizontal center for both datasets as illustrated in Fig. 3.

Fig. 3 Cropping the image.

For evaluation metrics, we use precision (P) and F-score by using the equation (1) and (2), respectively. In these equations, TP (true positive) is used for the number of correctly segmented pixels; FP (false positive) is used for the number of incorrectly segmented pixels; FN (false negative) is used for the number of pixels that are not segmented, but should be segmented; R is used for the recall.

$$P=\frac{TP}{TP+FP} \tag{1}$$

$$F\text{-}score=2*\frac{Precision*Recall}{Precision+Recall} \tag{2}$$

$$R=\frac{TP}{TP+FN} \tag{3}$$

A. AU-TSL Dataset Experiments

We randomly selected 1600 frames from our AU-TSL dataset videos. In order to first roughly annotate the ground truth regions in the selected images, we implemented a computer vision based algorithm. The algorithm uses the coordinates of the head and hands by using skeletal data to roughly draw circles around these regions. The background in these circles is eliminated by using the depth data. Therefore, if the hands do not intersect with body, they have more detailed ground truth as in Fig. 4a, if hands are in front of the body they have wider regions as in the Fig. 4b, 4c. We create the ground truth of our AU-TSL dataset by using this application.

We evaluated our model using two different configurations where the number of training frames differ, i.e. 400 and 200 frames. In the first experiment, we use only 400 frames for training and 200 frames for validation. The rest of 1000 frames are used as the test data. For the second experiment, we even reduce the training data to only 200 frames and validation data to 100 frames. We use 1000 frames

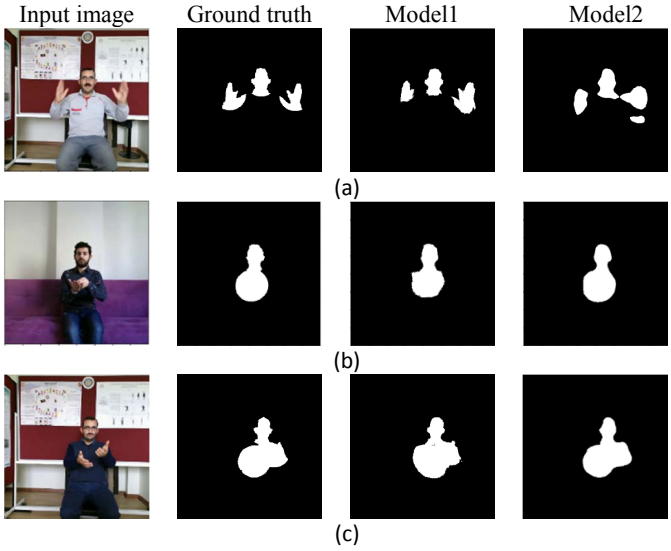

Fig. 4 Some segmentation results on our AU-TSL dataset. Some examples have more detailed ground truth for hands (a), while some have more rounded (b-c).

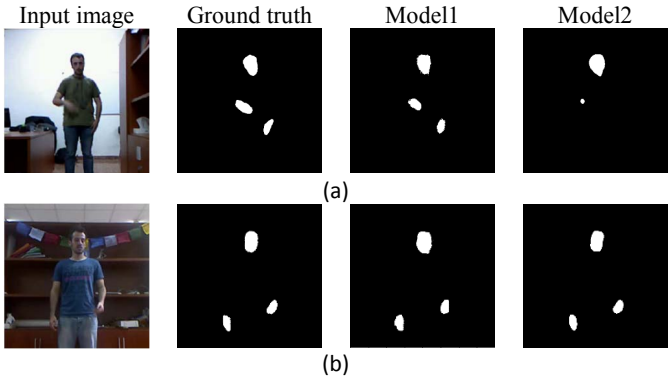

Fig. 5 Some segmentation results on Montalbano gesture dataset. Examples contains background and illumination variations, different signers and in some examples a) blurry hands, b) complex background.

as the test data. The results on AU-TSL dataset are reported in Table I.

B. Montalbano Dataset

We randomly selected 500 frames from Montalbano gesture dataset [8]. We manually mark the face and hand regions for ground truth regions generation. Same as the first dataset, we evaluate our model by using 400 and 200 frames for training. In the first experiment, where we use 400 frames for training, we use 33 frames for validation, 67 frames for test data. For the second experiment, where we use 200 frames for training, we use 100 frames for validation, 200 frames for test data. The results on Montalbano dataset are reported in Table II.

TABLE I. SEGMENTATION PERFORMANCE ON AU-TSL DATASET

	200 frames		**400 frames**	
	P (%)	F-score (%)	P (%)	F-score (%)
Model1	98	87.71	98.52	88.64
Model2	97.94	83.74	98.25	85.52

TABLE II. SEGMENTATION PERFORMANCE ON MONTALBANO DATASET

	200 frames		**400 frames**	
	P (%)	F-score (%)	P (%)	F-score (%)
Model1	97.67	76.34	98.78	80.2
Model2	98.53	75.16	98.59	76

C. Discussion

As seen in Table 1 and Table 2, our precision rates are quite high for both models in both datasets and there is not much difference in precision rates between the two models. F-scores are relatively lower since we have lower recall rates. Although Model1 (U-Net based) has fewer learned parameters, i.e. around 3.5 Million, than Model2 i.e. around 7 Million, it achieves better results than our second model (VGG based) when we compare their F-scores; it is because the first model has fewer false negatives. However, both models achieve comparable results inspite of using very few labeled training data. When we reduce the number of training data from 400 images to 200, precision rates do not change considerably; precision rates drop less than 1.2% on both models with both datasets. The variance in the F-scores is comparably higher, mainly due to the differences in the ground truth image generation on both datasets; in the worst case, where Model1 is run with Montalbano dataset, F-score drops 3.86%.

Some segmentation results are illustrated in Fig. 4 and Fig. 5. While Model1 is able to find fine details better, Model2 results are more rounded (see Fig. 4a). We believe that the primary reason is better domain adaptation; we trained Model1 from scratch, yet the encoder part of the Model2 is not adapted to the dataset.

Since we randomly select frames from videos, we do not manually control the selected images in the training and testing; hands can be blurry when they are moving. Therefore, in such cases, the models may fail to segment both hands (Fig. 5a), where hands look blurry in the frames.

As we mentioned before, while we use our annotation algorithm for the rough generation of the ground truth of our AU-TSL dataset, we manually marked the Montalbano dataset. Therefore, in AU-TSL, the annotated regions in the ground truth are larger, as can be seen in the Figure 4 and 5. F-scores are better in AU-TSL than Montalbano dataset. We believe that it is primarily due to the rough annotation, where ground truth frames do not have strictly marked boundaries in AU-TSL frames compared to the Montalbano frames. In the ground truths, segmented regions are larger and more rounded in AU-TSL, yet smaller and more accurate in Montalbano dataset. Therefore, we believe that our models can predict segmented regions more easily in AU-TSL dataset.

We develop our models on Tesla K80 from Google Colab. Training time of one epoch takes only 42 seconds in Model1, 110 seconds in Model2 when using 400 images for training. When using 200 images for training, training times of one epoch are 22 seconds and 60 seconds for Model1 and Model2, respectively. We terminated training when the validation accuracy stops improving. In test time, our models segment images with 400x400 resolution with 25 fps for Model1, 16.3 fps for Model2.

V. Conclusion

In this paper, we propose two encoder-decoder type deep learning based models for automatic segmentation of face and hands from static images. Our primary purpose is to generate a segmentation model that can be trained using only a few frames so that data annotation problem is reduced considerably for the datasets that are obtained from different environments. We experimented with two different datasets, where the ground truth images of each dataset are created differently, i.e. using a heuristic algorithm and using manual annotation by a person. Our experimental results show that the proposed networks works are well suited to face/hands segmentation problem, with few training data, i.e. 200 frames. Also, they are so fast; training takes approximately 42 minutes with 400 images and 22 minutes with 200 images. In test, average image segmentation is 25 frames per second with 400x400 resolution in Model1, 16.3 frames per second in Model2.

Acknowledgment

The research presented is part of a project funded by TÜBİTAK (The Scientific and Technological Research Council of Turkey) under grant number 217E022.

References

[1] Cheok, M. J., Omar, Z., Jaward, M. H. "A review of hand gesture and sign language recognition techniques", International Journal of Machine Learning and Cybernetics, 10(1), 131-153, 2019.

[2] Rautaray, S. S., Agrawal, A. "Vision based hand gesture recognition for human computer interaction: a survey", Artificial intelligence review, 43(1), 1-54, 2015.

[3] Krizhevsky, A., Sutskever, I., Hinton, G. E. 2012. "Imagenet classification with deep convolutional neural networks", In Advances in neural information processing systems, pp. 1097-1105.

[4] Long, J., Shelhamer, E., Darrell, T., "Fully Convolutional Networks for Semantic Segmentation", In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 3431-3440, 2015.

[5] Ma, M., Fan, H., Kitani, K. M., "Going deeper into first-person activity recognition", In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 1894-1903, 2016.

[6] Ronneberger, O., Fischer, P., Brox, T. "U-net: Convolutional networks for biomedical image segmentation", In International Conference on Medical image computing and computer-assisted intervention, 2015, October, pp. 234-241, Springer, Cham.

[7] K. Simonyan, A. Zisserman, "Very deep convolutional networks for large-scale image recognition", In Proc. International Conference on Learning Representations, arXiv:1409.1556, 2014.

[8] S. Escalera, X. Bar, J. Gonzlez, M.A. Bautista, M. Madadi, M. Reyes, V. Ponce, H.J. Escalante, J. Shotton and I. Guyon, "Chalearn looking at people challenge 2014: Dataset and results". In: ECCV workshop. 2014.

[9] Shukor, A. Z., Miskon, M. F., Jamaluddin, M. H., bin Ali, F., Asyraf, M. F., bin Bahar, M. B., "A new data glove approach for Malaysian sign language detection", Procedia Computer Science, 76, 60-67, 2015.

[10] Lamberti, L., Camastra, F. "Real-time hand gesture recognition using a color glove", In International Conference on Image Analysis and Processing, 2011, September, pp. 365-373, Springer, Berlin, Heidelberg.

[11] Shaik, K. B., Ganesan, P., Kalist, V., Sathish, B. S., Jenitha, J. M. M. "Comparative study of skin color detection and segmentation in HSV and YCbCr color space", Procedia Computer Science, 57, 41-48, 2015.

[12] Dardas, N. H., Georganas, N. D. "Real-time hand gesture detection and recognition using bag-of-features and support vector machine techniques", IEEE Transactions on Instrumentation and measurement, 60(11), 3592-3607, 2011.

[13] P. Viola and M. Jones, "Robust real-time object detection," International Journal of Computer Vision, vol. 2, no. 57, pp. 137–154, 2004.

[14] Huang, C.L., Jeng, S.H. "A Model-Based Hand Gesture Recognition System", Machine Vision and Application, 12(5), 243-258, 2001.

[15] Han, J., Awad, G., Sutherland, A. "Modelling and segmenting subunits for sign language recognition based on hand motion analysis", Pattern Recognition Letters, 30(6), 623-633, 2009.

[16] Desai, S. "Segmentation and Recognition of Fingers Using Microsoft Kinect", In Proceedings of International Conference on Communication and Networks, 2017, pp. 45-53, Springer, Singapore.

[17] Ren, Z., Yuan, J., Meng, J., Zhang, Z. "Robust part-based hand gesture recognition using kinect sensor", IEEE transactions on multimedia, 15(5), 1110-1120, 2013.

[18] Dadashzadeh, A., Targhi, A. T., Tahmasbi, M., Mirmehdi, M. "HGR-Net: A Fusion Network for Hand Gesture Segmentation and Recognition", arXiv preprint arXiv:1806.05653, 2018.

[19] Chen, L. C., Papandreou, G., Schroff, F., Adam, H. "Rethinking atrous convolution for semantic image segmentation". arXiv preprint arXiv:1706.05587, 2017.

[20] Matilainen, M., Sangi, P., Holappa, J., Silvén, O. "OUHANDS database for hand detection and pose recognition". In 2016 Sixth International Conference on Image Processing Theory, Tools and Applications (IPTA) (pp. 1-5), 2016, December, IEEE.

[21] Grzejszczak, T., Kawulok, M., Galuszka, A. Hand landmarks detection and localization in color images. Multimedia Tools and Applications, 75(23), 16363-16387, 2016.

[22] Shahriar, S., Siddiquee, A., Islam, T., Ghosh, A., Chakraborty, R., Khan, A. I., Shahnaz, C., Fattah, S. A. "Real-Time American Sign Language Recognition Using Skin Segmentation and Image Category Classification with Convolutional Neural Network and Deep Learning". In TENCON 2018-2018 IEEE Region 10 Conference (pp. 1168-1171),2018, October, IEEE.

[23] Neverova, N., Wolf, C., Taylor, G. W., and Nebout, F. "Multi-scale deep learning for gesture detection and localization". In Workshop at the European Conference on Computer Vision (pp. 474-490). Springer, 2014.

[24] Pigou, L., Dieleman, S., Kindermans, P. J. and Schrauwen, B. "Sign language recognition using convolutional neural networks", In Workshop at the European Conference on Computer Vision (pp. 572-578). Springer, Cham, 2014.

[25] Tompson, J., Goroshin, R., Jain, A., LeCun, Y., Bregler, C. "Efficient object localization using convolutional networks", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 648-656, 2015.

[26] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. "Dropout: a simple way to prevent neural networks from overfitting", The journal of machine learning research, 15(1), 1929-1958, 2014.

[27] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. "Imagenet large scale visual recognition challenge", International Journal of Computer Vision, vol. 115.3: pp. 211-252, 2015.

[28] Kingma, D. P. And Ba, J., "Adam: A method for stochastic optimization", arXiv: 1412.6980, 2014.