HAND SEGMENTATION FOR ARABIC SIGN LANGUAGE ALPHABET RECOGNITION

Ouiem Bchir

College of Computer and Information Sciences
Computer Science Department, King Saud University
Riyadh, Saudi Arabia

ABSTRACT

This research aims to separate the hands from the background of colored images representing the Arabic Sign language alphabet gestures. This hand segmentation task is one of the main challenges of image based Sign language recognition systems due to the issue of skin tones variations and the complexity of the background. For this purpose, an efficient system that segment the hand object and separate it from the rest of the image based on deep learning is investigated. More specifically, the DeepLab v3+ network architecture that is a combination of spatial pyramid pooling module and encode-decoder structure will be trained to learn the visual characteristics of the hand and segment it with detailed boundaries. The effectiveness of the proposed solution is investigated on a large dataset of size 12000 with an accuracy of 98%, an IoU of 93% of and BF score of 87%.

KEYWORDS

Hand segmentation, Deep learning, Arabic Sign Language Alphabet, spatial pyramid pooling.

1. INTRODUCTION

Sign language is an essential way of communication for hearing and speaking impaired persons. In fact, in addition to their speaking disability, they usually exhibit reading and writing skill deficiency. However, even though sign language allowed persons with this disability to communicate among them, few people without the hearing/speaking disability understand it. This excludes the persons with disability from communication yielding their isolation, and thus affects heavily their life and generates sentiments of loneliness and frustration. This communication issue can be alleviated by the translation of the gestures of the sign language to a text and verse versa, especially with the recent technological advancements that have helped the development of systems that support sign languages recognition [1] [2], and thus have yielded the apparition of several systems on sign language recognition [3]. These systems are either based on images or on sensors.

Sensor-based systems require the utilization of gadgets like gloves or captors that capture the location and the gesture shape of the hand [4]. The problem with these gadgets is that they are cumbersome and non-convenient for the user to ware or to use. On the other hand, image based approaches are non-intrusive [5]. It is an alternative solution that has the advantage to be available and natural since it does not constrain the hearing impaired person to use any gadget and can be deployed using smart devices using the availability of cameras on the portable devices.

David C. Wyld et al. (Eds): ISPR, ARIN, MLCL, NBIoT - 2020
pp. 01-09, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.100701

Typically, image-based systems separate the hand object from the other objects in the scene and recognize the corresponding gesture using image processing and machine learning techniques. More specifically, the captured image of the scene, is first segmented into two objects, the hand and the background. The hand object is then conveyed as input to the recognition system. However, appropriate features need to be extracted from the image in order to efficiently discriminate the hand from the background. This choice is challenging and is considered as an issue of image-based systems [6] due to the variability of the skin colors and the complexity of the backgrounds.

In this paper, the segmentation problem of the hand is addressed. For this purpose, an efficient system that segment the hand object and separate it from the rest of the image for Arabic sign language alphabet recognition is investigated. The proposed approach is based on deep learning. Namely, the DeepLab v3+ network architecture [7] will be trained to learn the visual characteristics of the hand and segment it.

2. RELATED WORKS

Hand segmentation is the first task of image based sign language recognition system. However, it is not a straightforward task due the skin color variations and the complexity of the backgrounds. An interesting characteristic used for segmenting the hand is the skin color, since it is not sensitive to scale, location, and orientation of the hand. For this reason, several approaches used the pixel values to construct the hand model by employing parametric models such as Gaussian Mixture Model (GMM), and non- parametric models such as histogram based methods. Nevertheless, these color based models are sensitive to illumination and skin tone variations. To alleviate this issue, the chrominance components are usually used instead of the pixel color neglecting this way the luminance components. Nevertheless, there are some challenges that still restrict the segmentation process which are complexity of the background, and the low quality of the image.

Existing Arabic Sign language recognition systems addressed the segmentation problem in different ways. In fact, some approaches ([8],[9],[10],[11]) circumvent the segmentation task by restraining the gesture images to have a uniform background. Other reported works ([12] [13]) use depth cameras, such as Kinect, and segment the hand as the nearest object to the camera. On the other hand, other approaches made the segmentation problem easier by using gadgets such as colored gloves [14]. However, these approaches ([12] [13] [14]) are inconvenient due to the unavailability and the cost of the sensor or the cumbersomeness of the gloves.

Hand segmentation for Arabic Sign language alphabet recognition based on skin pixel's characteristics have been reported in the literature. The authors in [15] convert the RGB input image to the YCbCr space and then use the skin profile to separate the hand. To further enhance the segmentation results, morphological dilation [16] is performed. Similarly, for the purpose of hand segmentation for Arabic Sign language, the authors in [17] convert the RGB input image to the YCbCr space. On the obtained color space, the Grey Level Co-occurrence Matrix (GLCM) [16] is computed. Using the GLCM values, the contrast, the correlation, the energy, and the local homogeneity are calculated and conveyed to a Multiple Layer Perceptron classifier [18] in order to recognize the skin. On the other hand, the authors in [19] segment the hand using the iterative thresholding algorithm [16] after de-noising the input image using the median filter. Although the approach in [12] use depth cameras and the hand is segmented as the closest object using the depth information, they also perform pixel segmentation to deal with complex backgrounds. For this purpose, they use the RGB ratio model. It is defined as in (1)

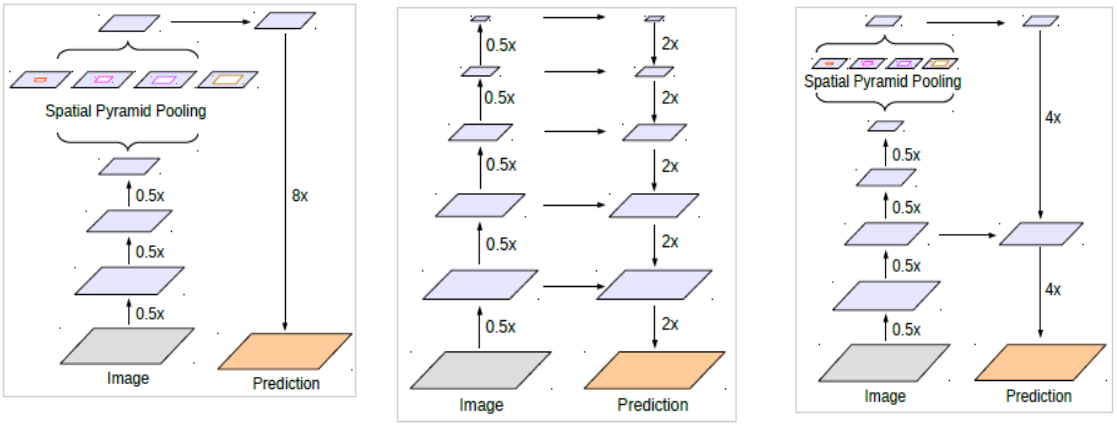
$$s = \begin{cases} 1 & if\ 0 \leq \frac{R-G}{R+G} \leq 0.5\ and\ \frac{R}{R+G} \leq 0.5 \\ 0 & otherwise \end{cases} \quad (1)$$

where s is the intensity of the segmented image, and R and G are the red and green color component of the input image.

As mentioned above, one of the main challenges of segmenting the hand is the choice of the suitable feature that allow discriminating the hand from the background. Deep learning based approaches can omit this choice since the feature is learned through the network architecture. However, Arabic Sign languages systems based on deep learning that have been reported so far ([20] [14] [13]) don't address the segmentation problem. They either consider hand images with uniform backgrounds, use depth sensors or gadgets. In summary, the hand segmentation for sign language recognition is still considered as an open issue problem.

Semantic segmentation that aims to predict a semantic label to every pixel [23] [25] is one of challenging tasks in computer vision. In this context of semantic segmentation, Fully Convolutional Networks (FCNs) [21,22] have shown promising results [23,24]. For this reason, various models have been proposed to extract the visual information [25, 26]. They include network models with multi-scale inputs [27, 28] or probabilistic graphical models like DenseCRF [17]. Besides, model networks such as PSPNet [29] or DeepLab [28, 7] that perform spatial pyramid pooling [30, 31] at several grid scales or apply atrous convolutions with different rates have demonstrated significant result improvement on several segmentation problems.

Recently, the DeepLabv3 model has been improved by combining the spatial pyramid pooling module (Figure1. (a)), with the encoder-decoder structure (Figure1. (b)) [32]. This results in the DeepLabv3+ model which learns rich semantic information from the encoder, and detailed object boundaries by the decoder. As shown in figure 1. (c), the encoder module extracts features at different resolutions by performing atrous convolution.



(a) (b) (c)

Figure 1. Combination of the spatial pyramid pooling module with the encoder-decoder structure, (a) Spatial Pyramid Pooling, (b) Encoder-Decoder, (c) Encoder-Decoder with Atrous Conv [32].

3. HAND SEGMENTATION SYSTEM

This research intends to segment the hand for Arabic sign language alphabet recognition using the DeepLabv3+ deep learning architecture. As in [32], two neural network modules are considered. Namely, spatial pyramid pooling module [33] and encoder-decoder structure [34] are used. The spatial pyramid pooling module is responsible for extracting the hand features and the encoder – decoder network learns the hand boundaries. In fact, even though the spatial pyramid pooling module encodes the visual information of the hand in the last layer, the hand boundary information is not available due to the pooling operation and the convolution striding through the network. A denser feature mapping can address the problem but it would be computationally prohibitive while the encoder-decoder network is fast. For this reason, as in [32], the two architectures are combined where the encoder module incorporates the feature extraction task. More specifically, we use the DeepLabv3+ architecture [32] that adds a decoder module that recovers the hand boundaries. Therefore, the extracted feature is encoded in the output of the spatial pyramid pooling module, and the hand boundaries are provided by the decoder module. The system is depicted in figure 2. As shown, it consists into two main components: the encoder and the decoder.

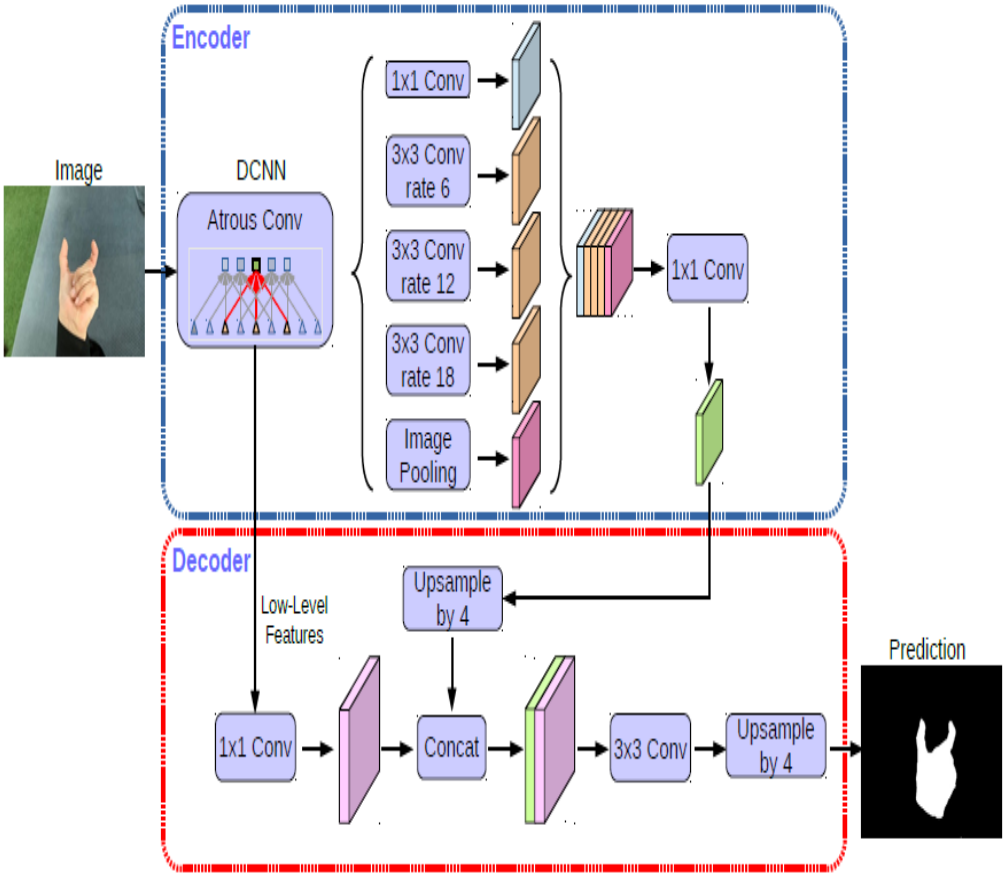


Figure 2. Segmentation system based on DeepLabv3+architecture [32].

The encoder module uses spatial pyramid pooling module [33] and employs atrous convolution [35] with different rates to extract the features. This results in a 256 coded vector that incorporate the semantic information of the hand at the output of the encoder. The decoder module up-

sample the output of the encoder bi-linearly by a factor of 4. The obtained vector is then concatenated to the convoluted low level feature. The concatenated feature is then convoluted with 3X3 filter and up-sampled by a factor of 4.

4. EXPERIMENTS

In order to assess the performance of the proposed system, a dataset of RGB images representing the 30 Arabic Sign Language gestures captured using mobile cameras is collected. Figure 3 displays samples from the dataset. The dataset includes 12000 images collected from 40 signers where each signer gestured the 30 letters in 10 different scenes. In order to construct the ground truth, the images were segmented manually where the hand is colored in white while the background is colored in black. The overall dataset is divided into 60% for training, 20% for validation, and 20% for testing.

Figure 3. Sample images representing the 30 gesture alphabets of the Arabic Sign Language alphabet.

The Deeplab v3+ network is setup using pre-trained Resnet-18 network for weight initialization. Table 1 reports the parameter's setting. Using the 60 % of the data and their corresponding ground truth images, the Network is trained. Then using the 20% of the validation data, the hyper-parameters are tuned. Finally, the system is tested using the remaining 20% of the data.

Table1. Parameter's setting

Learning rate	0.3
Learning period drop	10
Momentum	0.9
L2_Regularization	0.005
Max Epochs	30
Mini Batch Size	8
Shuffle	every epoch
Validation Patience	4

In order to evaluate the quality of the hand segmentation results against the ground truth, three performance measures are used, namely the accuracy, the Intersection over Union (IoU), also known as Jaccard similarity coefficient, and the BF Score. In order to compute these measure, the hand is considered as the positive class while the background is considered as the negative class. Thus, the number of true positives (TP) is the number of pixels belonging to the ground truth hand class and correctly segmented, and the number of false negative (FN) is the number of pixels belonging to the ground truth hand class and wrongly segmented as background. Similarly, the number of true negative (TN) is the number of pixels belonging to the ground truth background class and correctly segmented while the number of false negative (FP) is the number of pixels belonging to the ground truth background class and wrongly segmented as hand. The accuracy is then defined as in (2)

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

Since IoU is defined as the area of Intersection between the predicted hand and the hand in the ground truth divided over their union area, it can be expressed as in (3)

$$IoU = \frac{TP}{TP + FP + FN} \tag{3}$$

The BF Score which measures how close the learned hand boundary matches the ground truth hand is calculated as in (4)

$$BF = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

where

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

and,

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Using the three performance measures mentioned above, the quality of the hand segmentation results is evaluated against the ground truth. Table 2 reports the obtained results. As it can be seen, the segmentation approach performed very well on the considered Arabic sign language dataset. Figure 4 shows two examples that are correctly segmented.

Table 2. Performance results of the segmentation approach

Accuracy	IoU	BF Score
0.97941	0.92984	0.86562

In order to further investigate the obtained results, in Figure 5 hand segmentation examples where the segmentation was not correct are displayed. These bad result examples are related to the background characteristics. In fact, although the hand is detected, additional pixels from the background are wrongly predicted as hand. As it can be seen from Figure 5, this happens when two facts co-occur. The first one is that a part of the background color is similar to the hand skin color and the second one is that the overall region, constituted of the part of the background with similar color and the part of ground truth hand, has similar shape to another shape of the Arabic Sign language alphabet.

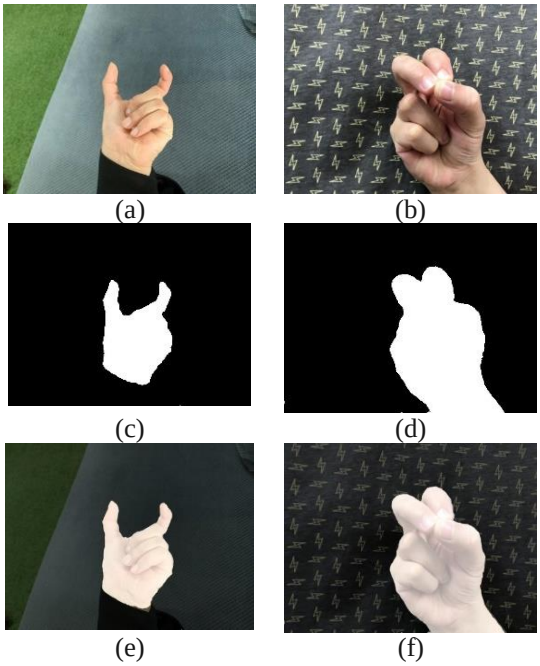


Figure 4. Two correctly segmented examples. (a) the first example of hand image, (b) the second example of the hand image, (c) the obtained segmentation of the image in (a), (d) the obtained segmentation of the image in (b), (e) the obtained segmentation over the original image in (a), (d) the obtained segmentation over the original image in (b).

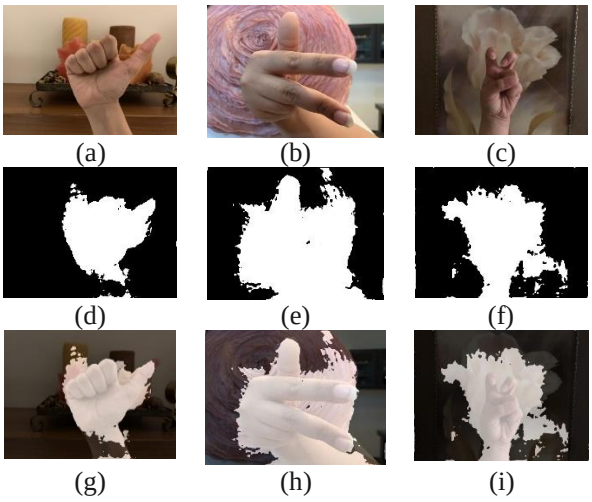


Figure 5. Three bad segmentation results. (a) the first example of hand image, (b) the second example of the hand image, (c) the third example of the hand image (d) the obtained segmentation of the image in (a), (e) the obtained segmentation of the image in (b), (f) the obtained segmentation of the image in (c), (g) the obtained segmentation over the original image in (a), (h) the obtained segmentation over the original image in (b), (i) the obtained segmentation over the original image in (c).

5. Conclusions and Future Works

In this paper, the DeepLabv3+ deep learning architecture is investigated for the segmentation of the hand pixels from images representing the Arabic Sign Language alphabets. The two modules of DeepLabv3+ allow effective semantic segmentation of the hand regions. In fact, the spatial pyramid pooling deep learning module foregoes the feature extraction stage as it is efficiently performed within the convolutional network. Moreover, the decoder module refines the segmentation results along the hand boundaries. The experimental results show the effectiveness of the proposed approach. As future works, the use of pre-trained networks other than Resnet-18 such as MobileNet v2 or ResNet-50 can be considered. Moreover, the obtained segmented images can be conveyed as input to an Arabic sign language alphabet recognition system.

References

[1] Tolba, M.F. and Elons, A. S., (2013) " Recent developments in sign language recognition systems", ICCES.

[2] Eisenstein, J., Ghandeharizadeh, S., Huang, L., Shahabi, C., G. Shanbhag, G., and Zimmermann, R., (2001) " Analysis of clustering techniques to detect hand signs", ISIMP.

[3] Brashear, H., Henderson, V., Park, K., , H., Lee, S. , and Starner, T., (2006) "American Sign Language Recognition in Game Development for Deaf Children", ASSETS, Portland, Oregon, USA.

[4] Mahesh, K., Mahishi, S., Pujari, S. R, N. S, and V., (2009) " Finger Detection for Sign Language Recognition", IMECS, Hong Kong

[5] Kang, S. K., Chung, K. Y., Rim, K. W., and Lee, J. H., (2011) " Development of Real-Time Gesture Recognition System Using Visual Interaction", IT Convergence and Security. Lecture Notes in Electrical Engineering, vol 120. Springer, Dordrecht

[6] Suharjito, F., Wiryana, F., Kusuma, G. P., and Zahra, A., (2018) " Feature Extraction Methods in Sign Language Recognition System: A Literature Review", INAPR .

[7] Chen, L.C., Papandreou, G., Schroff, F., Adam, H., (2017) " Rethinking atrous convolution for semantic image segmentation", arXiv.

[8] El-Bendary, N., Zawbaa, H. M., Daoud, M.S., Hassanien, A. E., and Nakamatsu, K., (2010) "ArSLAT: Arabic Sign Language Alphabets Translator",CISIM, Krackow, Poland.

[9] Sadeddine, K., Djeradi, R., Chelali, F. Z. and Djeradi, A., (2018) " Recognition of Static Hand Gesture", ICMCS.

[10] Tharwat, A., Gaber, T., Hassanien, A. E., Shahin, M. K., and Refaat, B., (2015) " SIFT-Based Arabic Sign Language Recognition System", Afro-European Conference for Industrial Advancement, Cham, 2015, pp. 359–370.
[11] Alzohairi R., Alghonaim, R., Alshehri, W., Aloqeely, S., and Bchir, O., (2018) " Image based Arabic Sign Language Recognition System", IJACSA, vol. 9, no. 3, 2018.
[12] Hamed, A. Belal, N. A. and Mahar K. M., (2016) "Arabic Sign Language Alphabet Recognition Based on HOG-PCA Using Microsoft Kinect in Complex Backgrounds",IACC, pp. 451–458.
[13] Aly, S., Osman, B., Aly, W. , and Saber M., (2016) " Arabic sign language fingerspelling recognition from depth and intensity images", ICENCO, Cairo, Egypt, 2016, pp. 99–104.
[14] Maraqa M. and Abu-Zaiter R., (2008) " Recognition of Arabic Sign Language (ArSL) using recurrent neural networks", ICADIWT, pp. 478–481.
[15] Hemayed E. E., and Hassanien, A. S., (2010) "Edge-based recognizer for Arabic sign language alphabet (ArS2V-Arabic sign to voice) ", ICENCO, pp. 121–127.
[16] Velho, L., Frery, A. C., and Gomes, J., (2009) ", Image Processing for Computer Graphics and Vision. Springer Science & Business Media.
[17] Dahmani, D., and Larabi, S., (2014) " User-independent system for sign language finger spelling recognition", Journal of Visual Communication and Image Representation, vol. 25, no. 5, pp. 1240–1250.
[18] Rosenblatt, F. F., (1963) "Principles Of Neurodynamics", Perceptrons And The Theory Of Brain Mechanisms.
[19] Al-Jarrah O. and Halawani, A., (2001) " Recognition of gestures in Arabic sign language using neuro-fuzzy systems", Artificial Intelligence, vol. 133, no. 1, pp. 117–138, Dec. 2001.
[20] Hayani, S., Benaddy, M., El Meslouhi, O. and Kardouchi, M., (2019) " Arab Sign language Recognition with Convolutional Neural Networks", ICCSRE, Agadir, Morocco, pp. 1–4.
[21] Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y. Overfeat. (2014) "Integrated recognition, localization and detection using convolutional networks",ICLR, 2014.
[22] Long, J., Shelhamer, E., Darrell, T., (2015) "Fully convolutional networks for semantic segmentation. CVPR, 2015.
[23] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., (2017) "Scene parsingthrough ade20k dataset", CVPR.
[24] Caesar, H., Uijlings, J., Ferrari, V., (2018) "COCO-Stuff: Thing and stuff classes in context", CVPR.
[25] Mostajabi, M., Yadollahpour, P., Shakhnarovich, G., (2015) "Feedforward semantic segmentation with zoom-out features", CVPR.
[26] Dai, J., He, K., Sun, J., (2015) " Convolutional feature masking for joint object and stuff segmentation", CVPR.
[27] Farabet, C., Couprie, C., Najman, L., LeCun, Y., (2013) "Learning hierarchical features for scene labeling", PAMI.
[28] Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., (2017) " Deeplab:Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs",TPAMI.
[29] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., (2017) "Pyramid scene parsing network", CVPR.
[30] Grauman, K., Darrell, T. (2005) " The pyramid match kernel, Discriminative classification with sets of image features",ICCV.
[31] Lazebnik, S., Schmid, C., Ponce, J., (2006) "Beyond bags of features, Spatial pyramid matching for recognizing natural scene categories", CVPR.
[32] Chen, L. C. et al., (2018) "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation", ECC.
[33] He, K., Zhang, X., Ren, S., Sun, J. ,(2014) "Spatial pyramid pooling in deep convolutional networks for visual recognition", ECCV.
[34] Badrinarayanan, V., Kendall, A., Cipolla, R. Segnet., (2017) "A deep convolutional encoder-decoder architecture for image segmentation",PAMI.
[35] Papandreou, G., Kokkinos, I., Savalle, P.A., (2015) "Modeling local and global deformations in deep learning Epitomic convolution, multiple instance learning, and sliding window detection", CVPR.