 doi:10.20944/preprints201811.0314.v1

Article

ROI Detection with Machine Learning for Glottis Images Captured from High-speed Video-endoscopy

Jun Kubota [1,2], Randol Spaulding [1], Chi Zhu [2], Krzysztof Izdebski [3] and Yuling Yan [1*]

[1] Department of Bioengineering, Santa Clara University, 500 El Camino Real, Santa Clara, CA, 95053, USA;
[2] Department of Systems Life Engineering, Maebashi Institute of Technology 460-1, Kamisadori, Maebashi-City, Gunma, Japan;
[3] Pacific Voice and Speech Foundation
* Correspondence: yyan1@scu.edu

Abstract: Detection of the region of interest (ROI) is a critical step in laryngeal image analysis for the delineation of glottis contour. The process can improve both computational efficiency and accuracy of the image segmentation task, which will facilitate subsequent analysis and characterization of the vocal fold vibration as it correlates with voice quality and pathology. This study aims to develop machine learning based approaches for automatic detection of ROI for glottis image sequences captured by high-speed video-endoscopy (HSV), a clinical laryngeal imaging modality. In particular, we first applied the supporting vector machine (SVM) method using histogram of oriented gradients (HOG) feature descriptor, and second, trained a convolutional neural network (CNN) model for this task. Comparisons are made for both approaches in terms of accuracy of recognition and computation time.

Keywords: High-speed video-endoscopy, laryngeal image processing, glottis delineation, Machine Learning, CNN

1. Introduction

Laryngeal imaging of voice production and subsequent image-based analysis of vocal fold vibrations are essential components of approaches to understand the mechanism of phonation and develop quantitative tools for the assessment of voice disorders[1,2]. The vast amount of video recordings produced from the HSV need to be processed to deliver useful, clinically relevant information. The ROI detection is important for the glottis image segmentation task since the process can significantly reduce computational cost while improve the accuracy of the segmentation results, key to the success of subsequent characterizations of the vocal fold vibration as it correlates with voice quality and pathology.

Vocal folds are deformable and the glottis images captured may differ in scale depending on both the HSV specifications and the examination procedure, and furthermore, the non-uniform illumination from the high-speed camera is expected that will fluctuate the image intensity, thereby imposing a challenge to the effective detection of the ROI.

Several issues are important in the detection of ROIs for glottis images, first being scale invariance. Our classifier model needs to be compatible with various datasets acquired from HSV systems with output images of different resolutions. Indeed, our datasets contained recordings from three different cameras with spatial image size of 512 by 512, 256 by 120, and 140 by 120 respectively. (Figure.1).

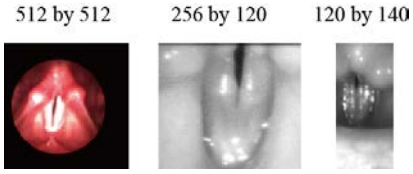


Figure 1. Representative image datasets with different pixel sizes.

Moreover, the scale of the vocal folds depends on how the examiner captures the images even with the same HSV system.

The second issue is brightness; several recordings exhibited strong light exposure, and the intensity variation among the recordings caused by non-uniform illumination is noticeable. Figure 2 shows two image frames selected from HSV recordings of two different patients where one of them (left) is overexposed to light. Such brightness difference may complicate the recognition task. For example,

histogram is widely used to describe feature characteristics that helps us understand color distributions of the objects, but it may not work in such environment. Thus, we must generate an ROI recognition model with robustness against brightness variation.

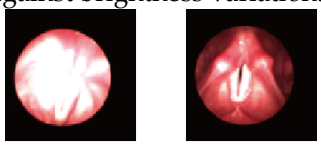


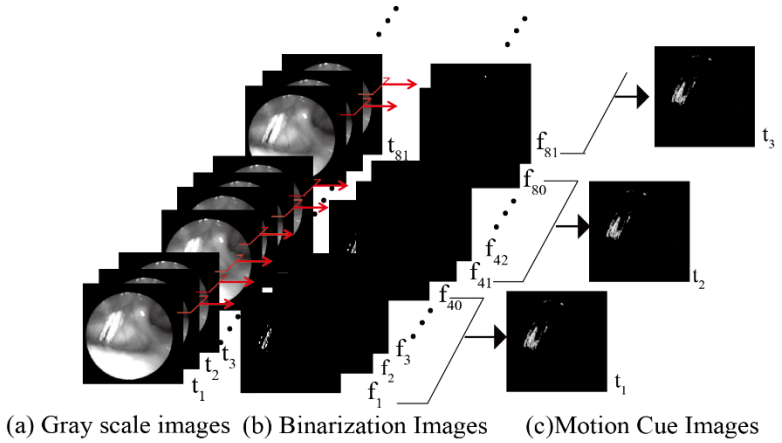
Figure 2. Different Brightness.

The third issue relates to the vocal fold deformation. While successful applications in broad areas of object recognition have been reported, the detection of the glottis region has not been developed. The fact that the vocal folds are deformable presents a challenge to the detection task.

To address the above mentioned issues, we propose a preprocessing step to generate new image sequences from the original recording based on motion cues. The preprocessed image sequences are then used to train both the SVM and the CNN classifiers.

2. Image Preprocessing

The new image sequences are binary and are constructed by first subtracting an image frame at time t from a frame at time t+1 (Figure 3.(a)), and then applying Otsu's thresholding method (Figure 3.(b)) so that the new binary image sequences are not influenced by brightness[3]. The new binary image has pixel values of 1 or 0, representing the region "in motion" or "static" respectively. Next, we generate a gray-scale image sequence containing motion cue, in particular, reflecting the range of the motion, by averaging each image with the subsequent images (up to 40 frames) (Figure 3.(c)). Figure 4 shows representative frames from five original video recordings (top) as well as the motion cue images generated (bottom). The original images were captured with an acquisition rate of 8000 fps (for first three) and 4000 fps (for the remaining two) respectively, and with various spatial resolutions (512 by 512, 256 by 120 and 120 by 140). The motion cue image sequences take into consideration both the deformation and the shape, thereby making the recognition task easier. Some images have noise, however, we can confirm that the movement of vocal folds is expressed in one image.



(a) Gray scale images (b) Binarization Images (c)Motion Cue Images

Figure 3. Motion Cue Images.

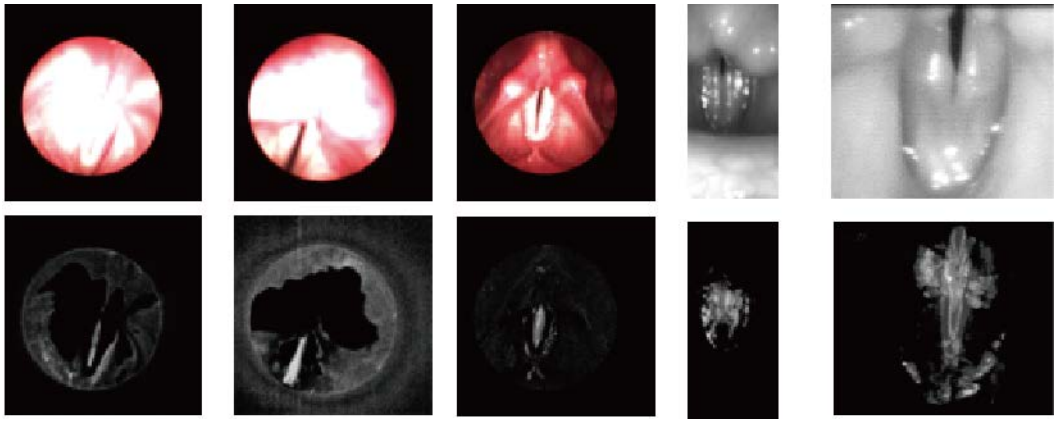


Figure 4. Representative image frames from original recordings and the corresponding motion cue images generated.

3. Detection of ROI using Machine Learning

After the pre-processing, we will perform the ROI recognition task using the motion cue images. Generally, there are two representative machine learning approaches for the object recognition: histogram of oriented gradient (HOG) feature descriptor, and convolutional neural networks (CNNs)[4,5]. Both methods have attracted attention from many researchers, and have shown high performance in terms of recognition accuracy and computation time. Here, we will explore which method is more appropriate for our application.

3.1. Histogram of Oriented Gradient (HOG)

HOG is one of the most representative feature descriptors in the field of image processing, and it is generally used for pedestrian detection.

Here, we provide a brief description of HOG and its application. As illustrated in Figure 5, the HOG calculates gradients and magnitudes in two filtered images (with X-gradient and Y-gradient filters).

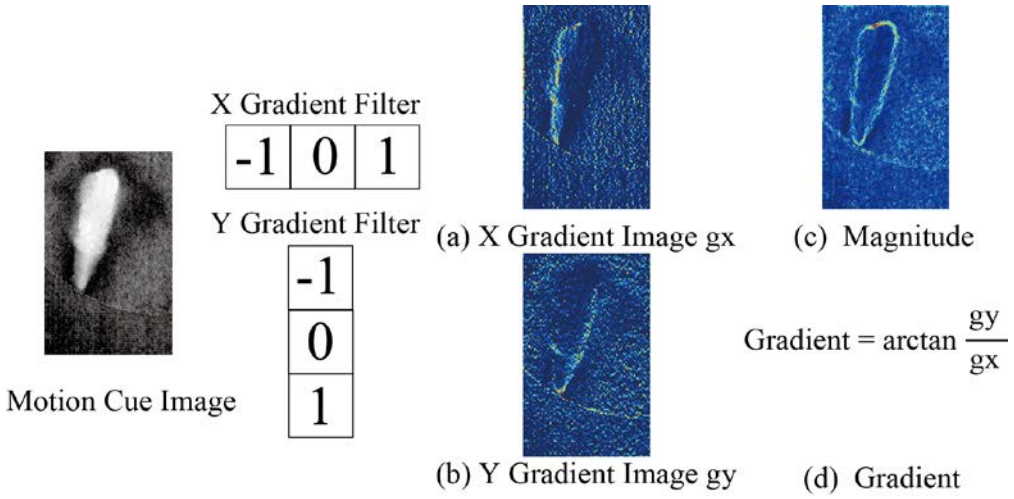


Figure 5. Magnitude and Gradient calculations in HOG.

The X gradient image (g_x) and Y gradient image (g_y) are expressed as follows:

$$g_x(i,j) = I(i,j) * f \quad (1)$$

$$g_y(i,j) = I(i,j) * f^t \quad (2)$$

 doi:10.20944/preprints201811.0314.v1

Where, I is the input image, and f = [-1,0,1] is the kernel. g_x and g_y emphasize the vertical and horizontal lines in the image respectively.

Using the two filtered images, the Magnitude and Gradient images are constructed as follows:

$$\text{Magnitude}(i,j) = \sqrt{g_x(i,j)^2 + g_y(i,j)^2} \quad (3)$$

$$Gradient(i,j) = arctan\frac{g_y}{g_x} \quad (4)$$

Next, we generate the histogram of the local gradients using magnitude and gradient images by dividing an image into cells each of 8 by 8 pixels.

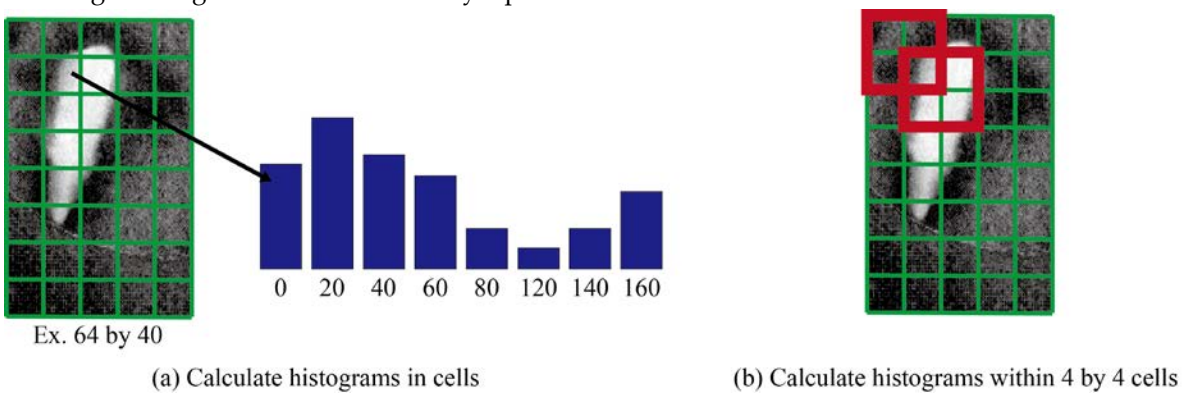


Figure 6. Calculate histograms

For instance, an image of pixel size 64 by 40 is divided into 40 cells of 8 by 8 pixels (Figure 6.(a)). We assign magnitude into a histogram for every angle (0 ~ 180 degrees) with an increment of 20 degrees. The calculated histograms for all cells are then normalized within the 4 by 4 cells (Figure 6.(b)). The HOG features consist of 1008 vector components for the image size of 64 by 40, which apparently contains information, however, two issues need to be considered.

First, the curse of dimensionality. It is known that recognition accuracy decreases as the number of feature vectors increase because significant information in higher dimension space for recognition is only a couple of components. Therefore, recognition accuracy may decrease because machine learning models consider redundant information if we apply high dimensional data.

Second, computational problem. To generate machine learning models such as SVM or K-NN, a number of vector components cannot be applied in terms of memory [6,7]. To address these issues, we applied Principle Components Analysis (PCA) to map HOG feature vectors into a lower dimensional space. Next, we generate a machine learning model to perform the recognition task in the reduced dimension. Many types of machine learning methods have been developed, and we chose SVM because of its favorable performance. Unlike K-NN or simple Neural Networks, SVM calculates distances from the function to the input data, and thus
the computation is relatively faster in comparisons.

3.2. Convolutional Neural Networks

Numerous studies using CNN in the field of image processing have been reported following the ImageNet Large Scale Visual Recognition Competition（ILSVRC） in 2012. Figure 7 shows the proposed CNN model consisting of two convolution and two max pooling layers and one fully connected layer for the ROI detection using the motion cue images as the inputs. The number of filters used for the two convolution layers is 32 and 64 respectively. We apply zero padding to both convolution layers, and a random dropout (50%) is implemented for the fully connected layer[8].

 doi:10.20944/preprints201811.0314.v1

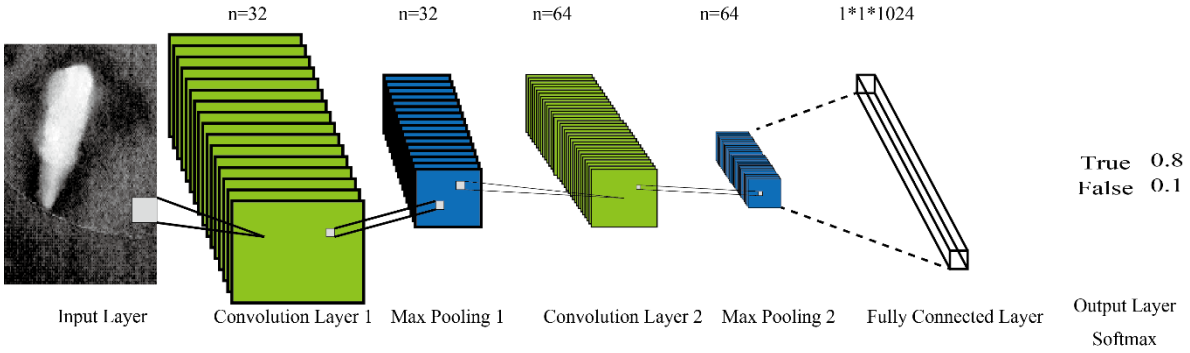


Figure 7. CNN used for ROI detection

4. Results

The true and false ROI datasets consisted of 15900 and 100 images respectively and were acquired from a variety of video recordings from male and female subjects with normal voice conditions or voice pathologies. Scales of vocal folds captured from video recordings are different each other. For instance, the vocal folds in the ROI dataset we used consists of from 40 by 25 up to 200 by 125. It would not be realistic to prepare machine learning models at each image sizes because it requires us to collect numerous dataset. Although resize processing may make images blur, we resized the ROI dataset that comprises the different image sizes to generate a machine learning model that is able to recognize the different scale of the ROI. We adopt a PC contribution ratio of over 80 and an RBF kernel for the SVM. The parameters of SVM were determined by eight cross-validation tests. The true and false ROI datasets are randomly selected (100 images from each category) for the testing. Tables 1&2 show the results of the recognition in terms of the accuracy and the computation time per image on average. Image size shows the resized size when we trained the models and recognized the test dataset. In terms of memory, we could not generate machine learning models with high resolutions. Therefore, we resized the ROI dataset up to 80 by 50.

Table 1. Recognition Accuracy[%].

Feature \ Image size	40 by 25	48 by 30	56 by 35	64 by 40	72 by 45	80 by 50
HOG	99.5	99.5	100	100	100	100
CNN	99.5	99.5	100	100	100	100

Table 2. Computation Time[sec].

Feature \ Image size	40 by 25	48 by 30	56 by 35	64 by 40	72 by 45	80 by 50
HOG	0.12	0.15	0.23	0.3	0.39	0.51
CNN	1.30	1.57	2.26	3.07	3.53	4.53

The recognition rate was shown to slightly decrease for the reduced image sizes (40 by 25 and 48 by 30). As the resized image enlarged, the recognition accuracy reached perfection and the calculation time increased accordingly.

5. ROI Identification

Following the ROI recognition with SVM or CNN performed on the motion cue image sequences, we need to select the optimal ROI from all possible regions that are recognized as true ROIs. This

 doi:10.20944/preprints201811.0314.v1

process is illustrated in Figure 8 where the recognized ROIs are mapped onto the original image sequences.

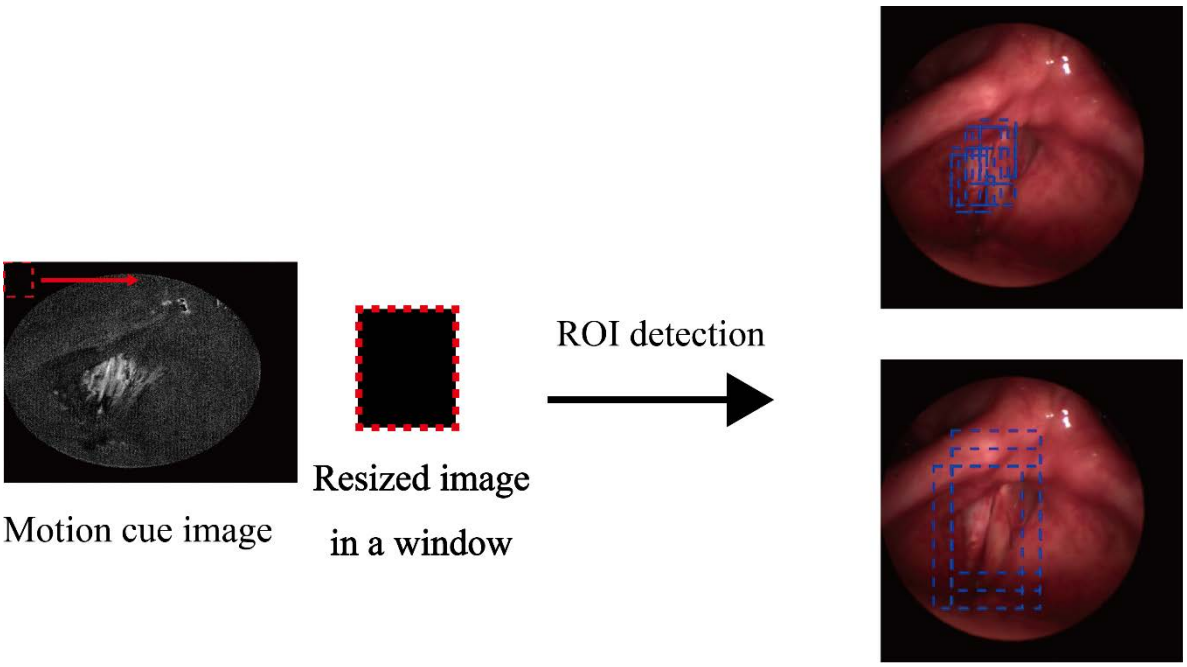


Figure 8. Processing for detecting ROI

For the recognition task using SVM or CNN, first, we slide windows by every 8 by 8 pixels, and an image inside the window is resized up to 64 by 40. The reason for limiting the window to this size is that the recognition accuracy no longer increases with further enlargement. The window is recognized as true ROI when it approaches the region where vocal fold motion occurs. We repeat the same process with a bigger window from the uppermost left until it reached the lowermost right of the image. This process recognizes ROI for multiple window sizes, so finally we need to determine the one window size and the coordinates. We determine the window size by choosing the median window size from the results. For example, we choose 80 by 50, if the windows of 64 by 40, 80 by 50, and 96 by 60 are all recognized as true ROI.

Once the window size is selected, we determine its location by calculating the average of the recognized coordinates. Figure 9 shows the detected ROIs using HOG (blue box) and CNN (red box) respectively. For comparisons, there is no significant difference between the two methods in terms of accuracy in recognition. Both correctly recognized the ROIs with high accuracy, although the HOG based recognition ran ten times faster than that using the CNN model.

5. Conclusion

For effective and practical ROI detection for the large glottis image datasets, we proposed to use the motion cue images for the recognition task. The ROI was detected using HOG based method or the CNN classifier. Both methods achieved high recognition accuracy, while the former requires much less computation time (about 10%) compared with the latter. This result would suggest that the HOG based method is a better choice for the ROI detection for glottis images considering the computational cost. Moving forward, we plan to further test the method on larger datasets. In the meantime, we will also explore applying deep CNNs to the laryngeal image-based classification of voice pathologies.

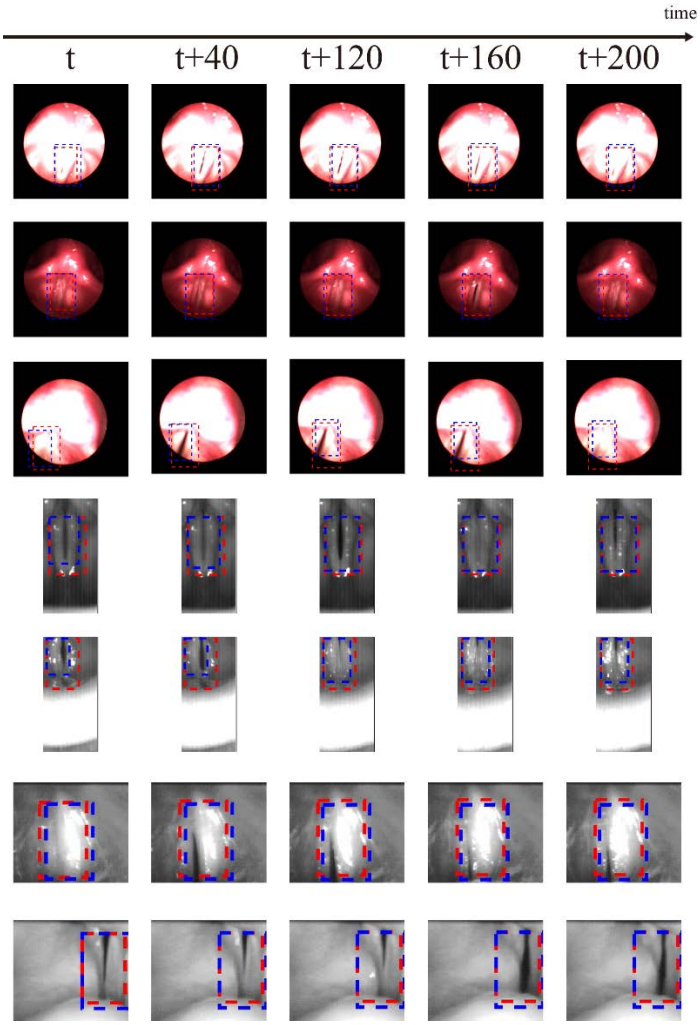


Figure 9. Result of detected ROI

References

1. Author 1, T. Shi.; Author 2, J. Kim, T. Murry.; Author 3, P. Woo.; Author 4, Y. Yan. Tracing Vocal-fold Vibrations Using Level-set Segmentation Method. Int. Journal for Numerical Methods in Biomedical Engineering, Vol. 31:6, 2015.
2. Author 1, Y. Yan.; Author 2, X. Chen.; Author 3 D. Bless. Automatic Tracing of the Vocal-fold Motion from High-speed Laryngeal Image Sequence. IEEE Trans. Biomedical Engineering 53: 7, pp1394-1400, 2005
3. Author 1, N.Othu. A Threshold Selection Method from Gray-Level Histograms. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 9, No. 1, pp. 62-66, 1979.
4. Author 1, N.Dalal.; Author 2, B.Triggs. Histograms of Oriented Gradients for Human Detection, Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp.886-893, 2005.
5. Author 1, A.Krizhevsky.; Author2, I.Sutskever.; Author3, G.Hinton. Imagenet classification with deep convolutional neuralnetworks. InProc. Of NIPS, 2012.
6. Author 1, I.Tsochantaridis.; Author 2, T.Joachims.; Author 3, T Joachims.; Author4, Y.Altun. Large Margin Methods for Structured and Interdependent Output Variables. Journal of Machine Learning Research 6, pp.1453–1484, 2005.
7. Author 1, B.Dasarathy. Nearest neighbor (NN) norms: nn pattern classification technique. IEEE Computer Society Press, 1991
8. Author 1, S.Nitish.; Author 2, G.Hinton.; Author 3, A.Krizhevsky.; Author 4, I.Sutskever.; Author 5, R.Salakhutdinov. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. Journal of Machine Learning Research 15 (1), pp1929–1958, 2015.