

# A Lightweight Feature Descriptor Using Directional Edge Maps

Wen-Hung Liao\*

**Keywords :** directional edge maps, local feature descriptor, object detection, robot vision.

## ABSTRACT

The objective of this research is to design a lightweight object detection and recognition engine that requires less space, less power and smaller budget than its PC counterparts. Specifically, we develop novel feature extraction algorithms to take advantage of fixed-point arithmetic. The newly defined descriptor, known as directional edge maps (DEM), can be computed using simple addition/subtraction operations. DEMs are employed as locally invariant features to represent objects of interest. When combined with a modified AdaBoost classifier, the system can be trained to detect and recognize objects of various types. The performance of the proposed descriptor in several object recognition problems are examined and compared in terms of accuracy and efficiency against local binary descriptors (LBP) and Haar-like features.

## INTRODUCTION

Designing a sociable robot that can communicate and interact with human in a natural manner requires a vision system that is robust and responsive. It is, however, a challenging task due to the variability of environments that brings about issues regarding lighting conditions, deformation, and occlusion. Rapid and accurate object recognition is therefore a subject that has been under continuous investigation in recent years when service-type robots become available.

Unlike previous robot vision system that focuses on topics such as scene reconstruction, autonomous navigation, obstacle avoidance, or path planning (Chatterjee *et al.* 2012), a service-type robot needs to know more about the environment. In other words, it needs to see and react to different objects in the surrounding. A sociable robot should be able to locate, recognize and track the user in real-time. It would help tremendously if gesture commands and facial expression can be understood (Breazeal 2004). In any case, the ability to recognize becomes a vital skill for any intelligent robot designed to interact with humans.

In this paper, we propose a novel feature descriptor: directional edge map (DEM) which possesses several desirable properties, including ease of computation, small memory footprint, robustness to noise, and the ability to describe objects with proper training. When combined with a modified AdaBoost classifier, the proposed scheme can perform object recognition tasks such as face detection, facial expression analysis and people counting efficiently and effectively. The claim is validated by experimental results using public databases as well as self-collected images.

The rest of this paper is organized as follows. In Section 2 we review related work regarding feature extraction and description for robot vision systems. Section 3 presents the local feature descriptor: directional edge map and the modified AdaBoost classifier employed in this research. Section 4 discussed experimental results of applying the proposed object recognition engine to multi-pose face detection and facial expression analysis. Comparative analysis against the well-known Haar-like method and local binary patterns is performed for the face detection task. Section 5 concludes this paper with a brief summary and outlook on future enhancements.

## RELATED WORK

Traditionally, robotic vision system relies on low-level features such as edges, corners or lines due to the need to process and analyze acquired images in real time. These relatively simple features may well serve the purpose of recovering structure of the scene or tracking motion of certain objects. However,

*Paper Received March, 2013. Revised July, 2013. Accepted December, 2013. Author for Correspondence: Wen-Hung Liao.*

\* Associate Professor, Department of Computer Science, National Chengchi University, Taipei, Taiwan 116, ROC.

when it comes to the problem of detecting or recognizing objects in the scene, more sophisticated ways of describing the objects are needed (Tuytelaars *et al.* 2008).

Scale-invariant feature transform (SIFT) (Lowe 2004) is a robust local feature descriptor that can deal with changes in lighting, viewing angle and deformation. It is the core of many image retrieval systems. However, its high computational cost discourages interactive robot applications.

The Haar-like features and cascade AdaBoost algorithm proposed in (Viola *et al.* 2004) has become a standard method to perform object detection. The set of Haar-like features is usually very large. The corresponding model stored in XML format therefore requires significant amount of memory. The situation becomes more critical when multiple models representing different poses or parts are loaded. The DEM feature set developed in this research is smaller, and can usually converge to the correct answer more quickly. Regarding the choice of classifier, AdaBoost algorithm is basically a two-category classifier. On the other hand, the modified AdaBoost is devised to address multi-class recognition problem.

Local binary patterns (Ojala *et al.* 2002) are computationally efficient descriptor that has found success in many object recognition applications. It is, however, quite sensitive to noise and has limited capabilities in handling regions where textures are not apparent. The proposed DEM follows a similar concept to partition the filtered response, yet with more intervals (5 instead of 2) to prevent the potential difficulties caused by binary thresholding.

Support vector machines have been extensively used for classification tasks with great success. Recently, ensemble of exemplar-SVMs have been applied to recognize object categories (Malisiewicz *et al.* 2011). The foregoing review focuses on feature descriptors. With proper modifications, DEM may serve as the input vector of the SVMs.

## FEATURE EXTRACTION AND CLASSIFICATION

In this section, we present and discuss the key components of the proposed object detection framework, namely, the descriptor: directional edge maps, and the classifier: modified AdaBoost. Formal definition regarding the local feature descriptor and the classifier will be given. Their important properties will then be elucidated.

### Directional Edge Maps

We propose a novel feature descriptor named directional edge maps (DEM) to facilitate the object detection and classification tasks. DEM has its root from edges and Haar-like features. Orientation information is preserved by using kernels that

extracts edges of different directions, as shown in Figure 1.

$$\begin{aligned}
 G_0 &= \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} & G_1 &= \begin{bmatrix} 0 & 0 & 0 \\ -1 & -1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \\
 G_2 &= \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \end{bmatrix} & G_3 &= \begin{bmatrix} 0 & -1 & -1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix} \\
 G_4 &= \begin{bmatrix} 1 & 1 & 1 & 0 \\ -1 & -1 & -1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} & G_5 &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\
 G_6 &= \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ -1 & -1 & -1 & 0 \\ 1 & 1 & 1 & 0 \end{bmatrix} & G_7 &= \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\
 G_8 &= \begin{bmatrix} 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} & G_9 &= \begin{bmatrix} 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} \\
 G_{10} &= \begin{bmatrix} 1 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 1 & -1 \\ 0 & 0 & 0 & 0 \end{bmatrix} & G_{11} &= \begin{bmatrix} 0 & 0 & -1 & 0 \\ 0 & -1 & 1 & 0 \\ -1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}
 \end{aligned}$$

Fig. 1. 12 kernels to define the directional edge maps

Twelve convolution kernels are employed to define DEM. Kernels  $G_0$  to  $G_3$  are of size  $3 \times 3$  and  $G_4$  to  $G_{11}$  are of size  $4 \times 4$ . The different kernel sizes introduce more variability, and thus, more descriptive ability of the feature. The coefficients in the kernels are either 1 or -1, making the filter response very fast to compute since only integer addition/subtraction is involved.

To obtain the DEM representation, an input image  $A$  will first be convolved with the kernels  $G_0$  to  $G_{11}$  to compute the responses  $R_0$  to  $R_{11}$ . The results are divided into five levels and encoded with numbers -2, -1, 0, 1, 2 according to Eq. (1), where  $\sigma$  stands for the standard deviation of  $R$  in the region, and  $\varepsilon$  is a scaling factor to adjust the interval. Such a formulation makes DEM adaptive to changes in lighting as the thresholds are not fixed in advance. In addition, the encoding method produces concise representation.

$$R_0(x, y) = \begin{bmatrix} 1 & 1 & 0 \\ -1 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} * A$$

$$D_0(x, y) = \begin{cases} 2, & R_0(x, y) > 2 \times \varepsilon \times \sigma \\ 1, & R_0(x, y) > \varepsilon \times \sigma, R_0(x, y) \leq 2 \times \varepsilon \times \sigma \\ 0, & |R_0(x, y)| \leq \varepsilon \times \sigma \\ -1, & R_0(x, y) < -\varepsilon \times \sigma, R_0(x, y) \geq -2 \times \varepsilon \times \sigma \\ -2, & R_0(x, y) < -2 \times \varepsilon \times \sigma \end{cases} \quad (1)$$

Figures 2(b)-(e) depict the DEM representation of a human face using kernels  $G_0$  to  $G_3$ . It can be observed that edges of different orientations have been captured and encoded succinctly using the proposed feature descriptor.

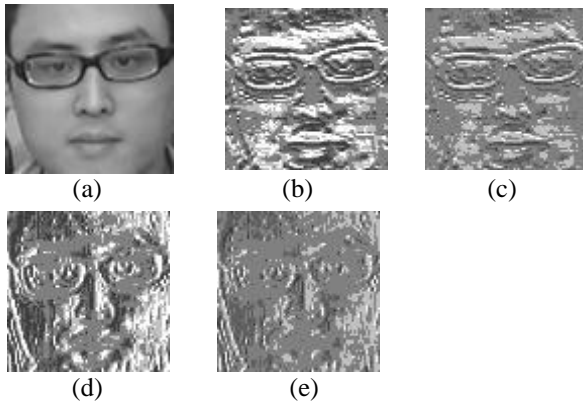


Fig. 2. DEM representation of (a) using kernels  $G_0$  to  $G_3$ .

In contrast to the Haar-like features and integral images adopted in (Viola *et al.* 2004), DEM attempts to capture the essence of an image region by a smaller, yet more representative set of features. This will effectively shorten the training process since DEM provides a higher level description of the scene.

The partition of the filter responses into five intervals and the subsequent encoding scheme reminds one of the thresholding process encountered in local binary patterns (LBP). However, the numbers of intervals are quite different in these two approaches. Moreover, DEM utilizes the encoded number directly whereas LBP requires another conversion process to obtain the final histogram representation. Since DEM is derived from image filtered with kernels containing only +1 or -1 as the coefficient, it is also invariant with respect to shift in image intensity, a desirable property that is also shared by LBP.

### Modified AdaBoost Algorithm

An object detection/recognition engine consists of at least two components: feature extraction and feature classification. We have modified the AdaBoost algorithm to cope with the classification problem in this research. The chief objective is to

handle multi-category classification with slight alterations of existing methods.

Given  $m$  samples  $(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$  where  $x_i$  denotes the input training data and  $y_i$  denotes the label of the sample (0 or 1).  $\omega_i$  specifies the weight for the  $i^{\text{th}}$  sample and is dynamically adjusted according to the current classification result. If a sample cannot be classified correctly, its weight will be increased so that it will have a higher probability of being selected for training in the next iteration. Originally there are  $p$  positive samples and  $n$  negative samples. Two output  $C(x)$  and  $Cg(x)$  will be generated after  $T$  iterations.  $C(x)$  represents the classification result and  $Cg(x)$  gives the confidence score. In contrast to the original AdaBoost algorithm where a binary decision is made, the proposed modification can handle multi-class recognition problems since we have the confidence score  $Cg(x)$  in our formulation. Steps of the modified AdaBoost algorithm are detailed in Figure 3.

1. Given  $(x_1, y_1), \dots, (x_m, y_m)$  where  $y_i \in \{0, 1\}$
2. Initialize  $\omega_i = \frac{1}{2p}$   $\omega_{i+p} = \frac{1}{2n}$

$p$ : number of positive samples     $n$ : number of negative samples

3. For  $t=1, \dots, T$ :
  4. – Train weak learner using distribution  $w_t$
  5. – Get weak hypothesis  $h_t : X \rightarrow \{0, 1\}$  with  $e_t = \min \sum_i \omega_i |h(x_i) - y_i|$
  6. – Choose  $\alpha_t = \frac{1}{2} \log \left( \frac{e_t}{1 - e_t} \right)$
  7. – Update:  $\omega_{t+1, i} = \frac{\omega_{t, i}}{\sum_j \omega_{t, j}} \exp(-\alpha_i)$
8.  $C(x) = 1, \text{ if } Cg(x) \geq \frac{1}{2} \sum_{i=1}^T \alpha_i$      $Cg(x) = \sum_{i=1}^T \alpha_i h_i(x)$

Output  $C(x)$  and  $Cg(x)$

Fig. 3. The modified AdaBoost algorithm for multi-category classification

### PERFORMANCE EVALUATION

We evaluate the performance of the proposed feature descriptor and classifier using two applications: multi-pose face detection and facial expression recognition. These two problems are important in designing intelligent robot-human interaction. It should be pointed out here that the proposed framework serves as a generic object recognition platform and can be trained to recognize other types of objects given proper training samples. As a matter of fact, the proposed methodology has been adapted to perform tasks such as license plate recognition and people counting.

### Multi-pose Face Detection

Face detection in video has received much attention due to its potential application in human-computer interaction. Although many techniques have been developed to resolve this issue, researchers are still in search for a robust method that can reliably detect all the faces in an image or a video regardless of head pose, viewing angle and lighting condition (Zhang *et al.* 2012).

The DEM-based approach is employed to detect human faces in multiple angles, as depicted in Fig. 4. To accomplish this task, we need to collect positive and negative samples, or face vs. non-face images. Since we are interested in multi-pose face detection, the positive training samples need to include faces of different orientations. For this experiment, we have prepared 20000 face images, i.e., 5000 for each of the four different angles: frontal, 30 degrees, 60 degrees and 90 degrees using collections from BaoDataBase, VidTIMIT, BioIDD and CMU-PIE database. 20000 non-face samples are collected from the Internet. All images are normalized to 24x26. Some examples of the training images are shown in Figs. 5-7.

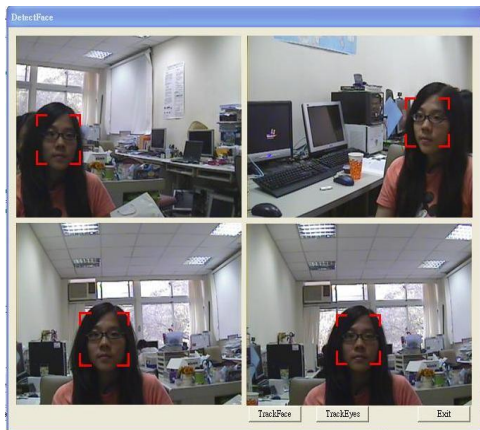


Fig. 4. Multi-pose face detection

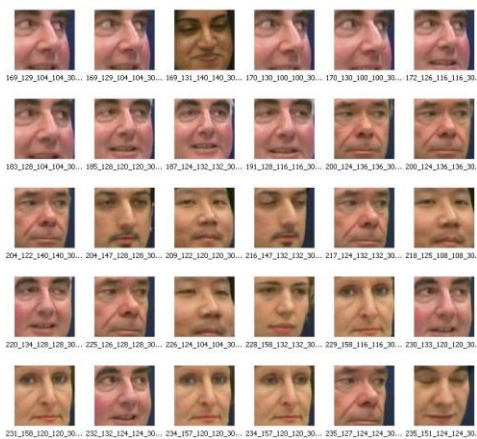


Fig. 5. Positive training samples (30 degrees)



Fig. 6. Positive training samples (90 degrees)

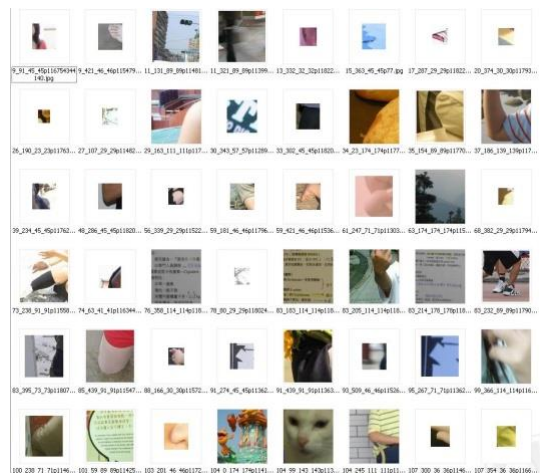


Fig. 7. Negative training samples

An experiment is conducted to compare the performance of the DEM-based method with the Viola and Jones algorithm using Carnegie Mellon University PIE face database (CMU). Results are shown in Fig. 8. The recall rate is approximately the same (79% vs. 81%). But the precision using our algorithm is much higher (90% vs. 77%).

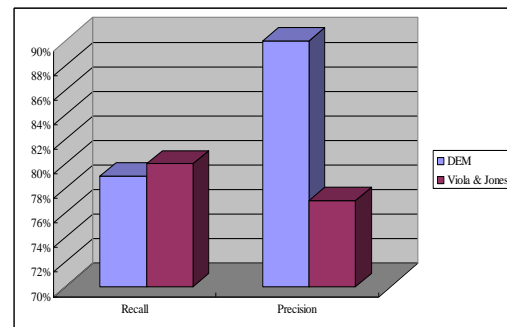


Fig. 8. Performance evaluation: precision and recall of DEM vs. Viola& Jones algorithm.

Fig. 9 compares the precision rate using DEM, LBP and Haar-like feature, respectively. The LBP model employed in this test is obtained from the latest OpenCV release (OpenCV 2013). The implementation is based on (Liao *et al.* 2007). The overall detection rate using LBP histogram is lowest among all three.

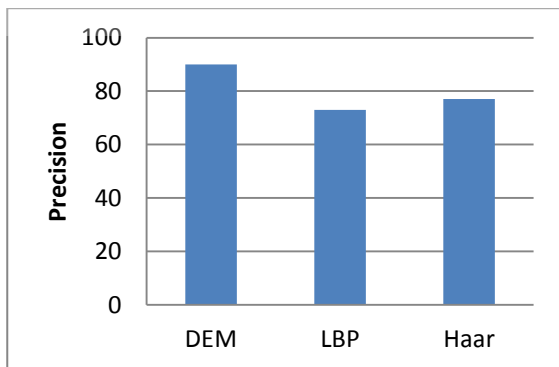


Fig. 9. Precision using DEM, LBP and Haar-like features

Regarding computational load, Table I lists the average processing time per image (size: 640x480) for different descriptors on a desktop PC with AMD9 945 CPU and 4G of RAM. The proposed method is most efficient, requiring 65% computation time on average when compared to Harr-like descriptors.

TABLE I: Computation time using different descriptors

Descriptor	Avg. computation time
DEM	50 ms
LBP	95 ms
Haar	77 ms

### Facial Expression Recognition

The ability to recognize emotion is critical in realizing affective computing (Picard 2000). A robot that can tell the emotion of the users from their facial expressions will be able to interact with the users in a delicate manner (Shan *et al.* 2009). Theoretically, facial expression analysis can be posed as a multi-class recognition problem that can readily be resolved using our proposed methodology. In practice, we find it necessary to combine rule-based information with the AdaBoost classifier to achieve best results, as shown in Figure 10.

The component-based model employs the rules listed in Table II to judge the tendency toward a certain type of emotion. Basically, we need to identify important facial components including the eyebrow, eye and mouth since their shape and deformation is directly associated with some types of emotion. Some results are shown in Figure 11. Notice that the information obtained here serves an

augmented role in the emotion recognition process and will be combined with the results from the action-unit model (Tian *et al.* 2001).

The action-unit model is essentially the DEM-based approach, with different sets of training samples to recognize different types of emotion. Since the resolution of the video is usually low (320x240), we will limit ourselves to the detection of salient changes in facial components, as illustrated in Table III. The results of applying the proposed method to facial expression recognition are summarized in Table IV.

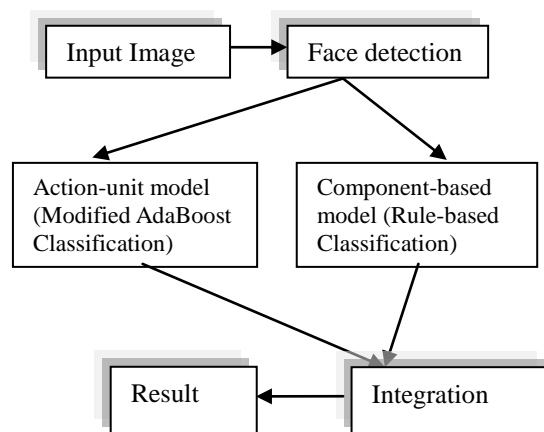


Fig. 10. Combining two models for facial expression recognition

TABLE II  
Rule-based Model for Emotion Recognition

Rule	Facial Component	Tendency
1	Eye closed	Sadness
2	Eyebrow up	Surprise
3	Mouth corner up	Happiness
4	Mouth corner down	Sadness/Anger
5	Mouth open	Happy/Surprise



Fig. 11. Extracting the mouth area for emotion recognition



TABLE III  
Action Units Employed in Our Model








	Forehead (wrinkle)
	Chin (wrinkle)
	Wrinkles on two sides of the face
	Mouth shape (smile)
	Mouth shape (anger)
	Mouth shape (surprise)
	Mouth shape (sadness)

TABLE IV  
Results of Emotion Classification

Type	Happiness	Sadness	Surprise	Anger
Result%				
Happiness	83.6	4.5	3.6	3.6
Sadness	5.5	72.7	8.2	11.8
Surprise	2.7	12.7	80	3.6
Anger	6.4	4.5	6.4	78.1
Neutral	1.8	5.5	1.8	2.7

## CONCLUSIONS

A novel and light-weight method for detecting/recognizing objects in real-time has been presented. The newly devised DEM feature possesses many desirable properties, including ease of computation, small memory footprint, and robustness to noise. Experiments have been conducted to validate the efficacy and efficiency of the proposed approach in multi-pose face detection and facial expression recognition tasks. Since DEM features can be computed using very simple operations, the proposed object detection/recognition engine can be efficiently implemented on mobile robots.

## REFERENCES

- Breazeal C., *Designing Sociable Robots*, A Bradford Book, (2004).
- Chatterjee A, Rakshit A., N. N. Singh N. N., *Vision Based Autonomous Robot Navigation: Algorithms and Implementations*, Springer, (2012).

CMU PIE Database:

[http://www.ri.cmu.edu/research\\_project\\_detail.html?project\\_id=418&menu\\_id=261](http://www.ri.cmu.edu/research_project_detail.html?project_id=418&menu_id=261)

- Liao S.C., Zhu X. X., Z. Lei, L. Zhang and Li Stan, "Learning multi-scale block local binary patterns for face recognition", *International Conference on Biometrics*, pp. 828-837, (2007).
- Lowe D., "Distinctive image features from scale-invariant keypoints", *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91-110, (2004).
- Malisiewicz T, Gupta A., and Efros A. A., "Ensemble of exemplar-SVMs for object detection and beyond", *International Conference on Computer Vision*, (2011).
- Ojala T. and Pietikainen M., "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns" *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 24, pp. 971-987, (2002).
- OpenCV 2.4.4 documentation (retrieved 3/17/2013): [http://docs.opencv.org/doc/tutorials/objdetect/cascade\\_classifier/cascade\\_classifier.html](http://docs.opencv.org/doc/tutorials/objdetect/cascade_classifier/cascade_classifier.html)
- Picard R. W., *Affective Computing*, The MIT Press, (2000).
- Shan C., Gong S., and McOwan P. W., "Facial expression recognition based on local binary patterns: a comprehensive study", *Image and Vision Computing*, Vol. 27, Issue 6, pp.803-816, (2009).
- Tian Y, T. Kanade T and J. F. Cohen J. F. , "Recognizing action units for facial expression analysis", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, Issue 2, pp.97-115, (2001).
- Tuytelaars T. and Mikolajczyk K., "Local invariant feature detectors: a survey", *Foundations and Trends in Computer Graphics and Vision*: Vol. 3: No 3, pp 177-280, (2008).
- Viola P. and Jones M., "Robust real-time face detection," *International Journal of Computer Vision*, Vol. 57, Issue 2, pp.137-154, (2004).
- Zhang C. and Zhang Z., "A survey of recent advances in face detection". Technical Report, Microsoft Research, (2010).