


# The Big Data Analysis on the Camera-based Face Image in Surveillance Cameras\*

Zhiguo Yan, Zheng Xu  and Jie Dai

Department of Internet of Things, The Third Research Institute of the Ministry of Public Security, Shanghai, China

## ABSTRACT

In the Big-Data era, currently how to automatically realize acquisition, refining and fast retrieval of the target information in a surveillance video has become an urgent demand in the public security video surveillance field. This paper proposes a new gun-dome camera cooperative system, which solves the above problem partly. The system adopts a master-slave static panorama-variable view dual-camera cooperative video-monitoring system. In this dual-camera system the gun camera static camera) with a wide viewing -angle lenses is in charge of the pedestrian detection and the dome camera can maneuver its focus and cradle orientation to get the clear and enlarged face images. In the proposed architecture, Deformable Part Model (DPM) method realizes real-time detection of pedestrians. The look-up table method is proved feasible in a dual-camera cooperative calibration procedure, while the depth information of the moving target changes slightly. As respect to the face detection, the deep learning architecture is exploited and proves its effectiveness. Moreover, we utilize the Haar-Like feature and LQV classifier to execute the frontal face image capture. The experimental results show the effectiveness and efficiency of the dual-camera system in close-up face image acquisition.

## KEYWORDS

Gun-dome camera cooperation; calibration; deformable part model; pedestrian detection; LBP

## 1. Introduction

According to statistics, more than 90% of criminal cases need access to a surveillance video system to achieve related evidence. However, the current applications of video data mainly are based on artificial operation supplemented by simple intelligent video analysis. There exist many problems impeding the use of a massive video, such as “exist but can’t find it”, “Takes too much time to find it”, “Services unreliable,” etc. In the upcoming Big-Data era, this embarrassment has become more emergent and unacceptable. Therefore, how to automatically realize acquisition, refining and fast retrieval of the target information in a video surveillance system has become an urgent demand in the field of public security video surveillance.

In the point-view of Big Data analysis, the acquisition of interested objects, such as the related people and vehicles in some important events, plays the fundamental role in data cleaning and data vitalization. By gathering the interested data and figuring out their identity label attribution, then combining the content-based analysis technique and semantic description technique, there generate the new generation efficient search engine framework for the mass video data in the video surveillance system.

At present, there are two common solutions to trace and capture images of concerned targets in a surveillance video. One is on the basis of continuous tracking of the moving target with a single dome camera, and the other bases on the concerned target collaborative tracing using omnidirectional camera and active camera. In the former solution, with the object of attention appears in the dome camera wide scene, it will be focused, zoomed, and traced continuously. Nevertheless, the other target information in the wide scene will be neglected. Because once a focusing target appears, only one dome camera fully zooms its view to the target area and starts to track the

target continuously. For the latter method, the omnidirectional camera transmits position information of the concerned object to the active camera, and the object was further consistently traced by the active camera. However, this kind of application mode currently does not have a broad application prospect. The reason is that the narrow application of the omnidirectional camera in the visual surveillance field and the much high requirement of dual-camera calibration techniques, the complicated operation and the high requirement of operators’ professional, which will be accounted by the inconsistency of the resolution from the centre to the periphery and nonlinear mapping of moving trajectory of the omnidirectional camera.

In this paper, we propose a new gun-dome camera cooperative system, which adopts master-slave static panorama-variable view dual cameras cooperative video-monitoring system. Compared with the above solutions, it has the following advantages: Combining the advantages of the static-panorama camera and the camera with variable field of view (FOV), we can get the close shot of specific objects in the long shot. Meanwhile, we can also keep the attention to others objects in the distant scenery. By using this mechanism, we can expand the breadth and depth of video surveillance system. FOV of panorama camera is large, and distant scenes can be observed via focusing the moving camera. Complementary advantages could be obtained by combining these two cameras. 2) Realize observation of multiple targets. If only the moving camera is used for observation of a target, the FOV of the camera becomes small. When other targets appear in the FOV, they could not be observed from the FOV. 3) Facilitate the detection of the target. Moving target detection method can be used to detect the target, since the wide angle camera is static. And if the target motion is not so rapid in the FOV of the panorama camera, it is easy to trace.

Deformable Part Model (DPM) is used as target detection method in this system. In this paper, the look-up table method is proven feasible when the depth of field of moving target changes small. It is used to calibrate the dual-camera system, and it can obtain the angle of rotation, which the moving camera rotates to aim at the arbitrary position of the static camera. In this study, we present a facial orientation recognition method based on LVQ. In which, the features vector is comprised based on the features detection such as the edge of eyes, which was processed by edge detection method such as method based on Haar-like feature.

The organization of this paper is as follows: The gun-dome cooperative camera system will be introduced in Section 1. Methodology will be introduced in Section 2. Section 3 elaborates the materials and the experiment setting. The experimental results and conclusions are given in Section 4.

## 2. Related work

Face detection refers to determining the presence and location of faces in an image. Face detection as the prerequisite plays an important role in face recognition task. Given an image frame obtained from real-time video surveillance system, the face detection task is to determine whether or not there exist any faces in the image. If exist, return the facial area location. In the past few years, many researchers have achieved some merit in indoor face detection, but as to the outdoor face detection, it is still a rough task (Felzenszwalb, McAllester, & Ramanan, 2008). The difficulty associated with face detection own to the following variable factors: Scale, face orientation, head pose, facial expression, illumination conditions, occlusions, etc. Skin Color, eye characteristic and Texture are the commonly used features in past decades. Due to the color offset of the surveillance camera and the shadow on the facial image, the detection performance is not well satisfied (Cho et al., 2012).

Deep Learning is the up-to-date achievement in artificial intelligence (Beriault, 2008). It mimics the deep architecture and the deep cognitive process of human brain and obtains the great success in classification tasks (Dong, Li, and Chen, 2009). As one kind of the artificial neural networks, CNN is also a method of deep learning. Unlike the traditional classification methods, convolution neural network (CNN) can learn features from input data automatically. Due to its special hierarchical structure, CNN shows strong robustness against geometric distortions, such as shifts, scaling, rotation, etc. Based on these merits, CNN can achieve good results in object detection and classification tasks. Therefore, CNN-based deep learning strategy was proposed for face detection in this paper.

Face orientation detection is the premise of face recognition. How to detect the face orientation and store the upright frontal face image for the further face recognition is of significance in civil video surveillance. Orientation is one of the basic characteristics in image understanding and pattern analysis. Many approaches have been proposed to solve the above problem. Li and Shen (2006) proposed orientation histogram for orientation analysis. Hao found that self-organizing fuzzy network with SVM (Hao, Zhang, and Yu, 2010) worked well in color image detection. Kim, Kim, and Song (2009) developed a good method to estimate the pose of a face, limited to in-depth rotations. By comparing matching in transparency displays, Xie, Dang, and Tong (2012) proposed an approach based on that for orientation tuning of human face. Xu and Chen (2015a), Xu, Liu, Mei, Hu, and Chen (2015b), Xu, Mei, Hu, and Liu (2016a),

Xu, Mei, Liu, Hu, and Chen (2016b) and Xu, Hu, and Mei (2016c) proposed a video structural description technology based framework. The proposed framework is used to detect the event in the surveillance videos.

Our approach is partly motivated by the previous work, in which they use the Haar-like feature and the boosting classification strategy to locate the eyes. On the basis of their achievement, we go a step further by using the Learning Vector Quantization (LVQ) classifier to detect the face orientation. To summarize the above-mentioned, we united the whole process of surveillance video, including pedestrian detection with dual-camera configuration, dual-camera calibration, face detection, frontal face image capture in the proposed scheme.

## 3. The dual-camera configuration

China's emergency operations follow similar settings used in other countries. Emergency managers at EOCs undertake the responsibility of overall coordination, command and control. They need to make decisions on a wide set of tasks, including situation assessment, action planning, resource allocation, personnel deployment and other relevant works. A variety of urgent and multifaceted tasks will then be distributed to multiple operation sectors and implemented in parallel. Decisions and actions should be based on a comprehensive understanding of disaster/incident contexts.

Gun-dome camera cooperative system is one kind of dual-camera monitoring system. There is one wide angle camera and one Pan Tilt Zoom (PTZ) dome machine. The wide angle camera is responsible for the target detection in wide field of view, and PTZ dome machine (also known as active camera) for focusing and amplifying and tracking continuously for the target of attention. Dual-camera cooperative system function mainly is composed of three parts; moving object detection in the wide-angle camera, calibration of the moving camera and the wide-angle camera, coordinated control of these two cameras.

The proposed gun-dome camera cooperative system is shown in Figure 1. It is key personnel target detection and recognition application platform architecture based on the gun-dome camera. In Figure 1, there is an overlapping region between the scene recorded by the gun camera and the scene recorded by the dome camera, which belongs to the joint calibration of overlapping scenarios. When calibrating, a dome camera is in the wide-angle state, and under this situation, the gun-dome machine has an overlapping region. The calibration objects should be placed in the overlapping region in the joint calibration. The gun camera is responsible for panoramic monitoring with a wide-angle shot. Moving target detection is focused on and target position information is transmitted to the dome camera under the panoramic field. Meanwhile, in the scene of a dome camera, according to the position mapping relationship obtained from the gun-dome camera cooperative calibration, we first transform the position of the moving target in the gun camera, and calculate the corresponding coordinate in initial scene of the dome camera, then we start real-time PTZ control and realize continuous tracing and facial image capture of the target of attention.

To realize a dual-camera cooperative system is needed? Observe and detect targets in the scene with static wide-angle camera; obtain the position of target and transmit the position information to the moving camera; moving camera tracing the target according to its position and amplify it. The

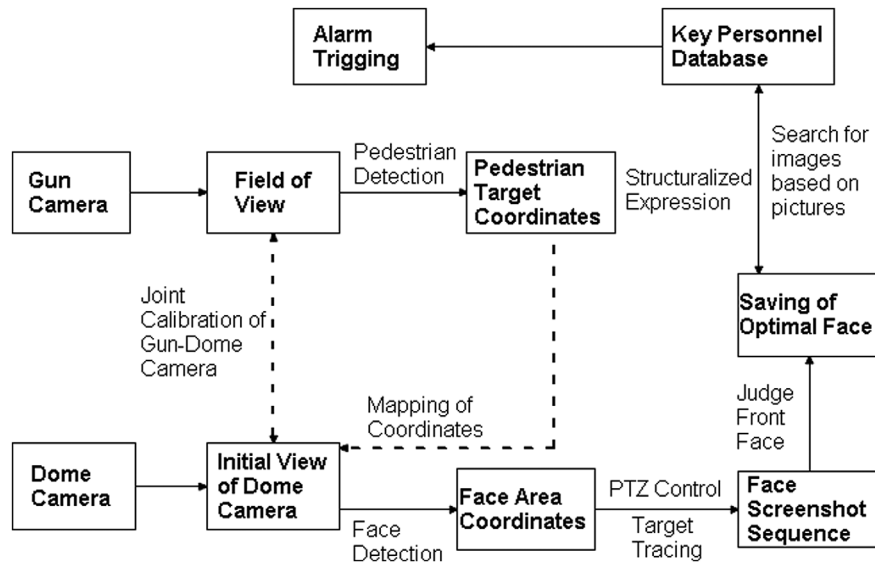


Figure 1. Block diagram of gun-dome camera cooperative system.

key techniques in this process is; (1) Target detection. Wide angle camera, which is static relative to the scene, can realize pedestrian objective location by motion detection. Moreover, illumination, swing of branches and any other factors should be taken into consideration in actual scene. (2) Calibration of dual cameras. The calibration of gun-dome camera is to compute the angle, which the moving camera rotates to aim at position of the target in the static camera. (3) Cooperative control of dual cameras. According to the state of the target in the static camera, appropriate cooperative strategy is set up and the clear image of the target is acquired.

## 4. Methodology

In this dual-camera system, DPM is used for target detection, Look-up Table Method is used to calibrate the gun-dome camera and CNN-based deep learning architecture and Haar-LQV are exploited to execute the face detection and orientation detection respectively.

### 4.1. Deformable part model (DPM)

DPM is a very successful target detection algorithm. It can be seen as an extension of Histograms of Oriented Gradients (HOG) and is consistent with HOG. First of all, HOG is calculated. Then, Support Vector Machine (SVM) is used for training to acquire the gradient model. These templates could be used for classification, which means matching the target with these templates.

Deformable part models such as pictorial structures provide an advisable framework for object detection. Yet it has been difficult to establish their value in practice, a single deformable model is often not expressive enough to represent a rich object category. Considering this problem of modelling the appearance of pedestrians in complicated scenes, we utilized the feature pyramid and the multi-resolution representation of the pedestrian model to improve the detection accuracy on small pixel-sized pedestrians normally missed by a single representation approach.

The standard procedure of DPM consists of these steps: Creating a densely sampled image pyramid, computing features at each scale, performing classification at all possible locations and finally performing non-maximal suppression to generate

the final set of bounding boxes. In these steps, the key influence factor includes the image pyramid and the computation intensity in search space and the feature extraction such as HOG.

When using DPM to execute the pedestrian detection, the former researchers generally limited the part number to four; head, leg, left arm and right arm. However, the optimal number of parts depends on the variability of an object class and may be significantly different between classes. In the practical experiments, we studied the optimal number of parts in pedestrian detection.

### 4.2. Calibration of dual cameras

The calibration of dual cameras is a process of computing parameters of geometric model of camera imaging. The Physical model method calculates the rotation angle based upon the imaging physical model of the target in the dual cameras and motion model of cameras. However, it can only obtain a very accurate rotation angle theoretically and the practical operation is much complicated. Compared with physical model method, Look-up table method is much convenient, simpler and reliable. In practice, the original calibration data is still valid when the scene changes from the learning environment to another.

In the gun-dome cooperative linkage personnel detection and tracking system, the gun camera has to transmit the position of pedestrian detected by itself to the initial scene of dome camera, and then based on this; the dome camera begins to do face detection and continuous tracking. The realization of the above functions requires of the mapping relationship between gun camera and dome camera. Namely the imaging position of the object in the gun camera is mapped into the dome camera, and the dome camera adjusts the rotation angle in order to make the object in the centre of image of the dome camera. Establishing this mapping relation is implemented through dual camera calibration. The calibration of the dual camera is referred to as; under the knowledge of the position of target in a wide angle camera, to find the horizontal rotated angle  $\alpha$  and the vertical rotated angle  $\beta$ , which make the PTZ camera rotate to aim at target,  $L[M(u, v)] \rightarrow (\alpha, \beta)$ .

The process can be elaborated in Figure 2. First of all, calibrate the internal parameters of the cameras; secondly eliminate the distortion of lens of cameras, then match the image

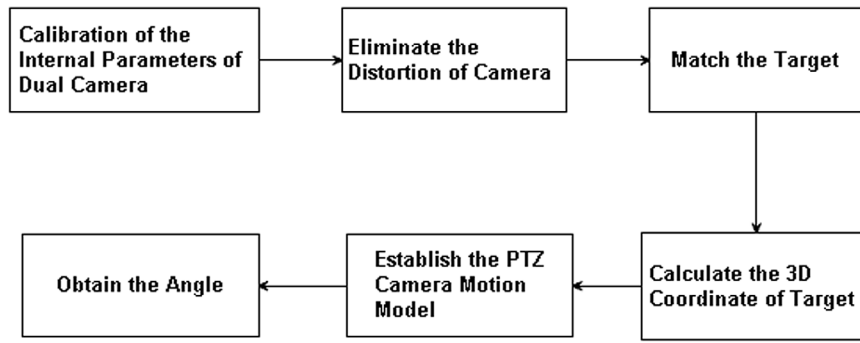


Figure 2. Process of dual camera calibration.

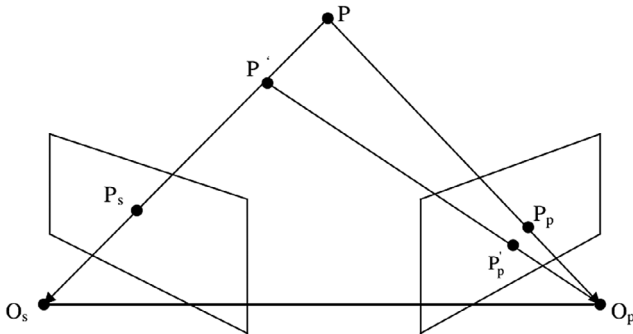


Figure 3. Dual cameras mapping.

target in dual camera via the polar constraint conditions and image features. Calculate the 3D coordinates of the target under PTZ camera coordinate system and finally compute the rotated angle  $(\alpha, \beta)$  according to PTZ camera motion model.

The imaging model of the dual cameras is shown in Figure 3. Through the analysis of the physical process of the dual cameras calibration, there exist a single mapping  $[x \ y \ z]^T \Leftrightarrow (\alpha, \beta)$  between the current position of the object  $[x, y, z]^T$  and absolute position parameter  $(\alpha, \beta)$  of camera when it is in the centre of the field of view of image. Thus, if the relationships between the three dimensional coordinate of the object  $[x, y, z]^T$  and the position of the camera  $(\alpha, \beta)$  is set up; the calibration of dual cameras will be realized.

The look-up table can be constructed by supervised learning, in order to acquire the angle, which PTZ camera rotates to aim at the target, and realize the dual cameras calibration. Detailed steps are as follows:

- (1) Choose Region of Interest (ROI), which needs PTZ camera's key monitoring in visual surveillance area of the wide angle camera.
- (2) Divide ROI of wide angle image into grids according to appropriate spacing. Then acquire the pixel coordinates in the grid intersections.  $M_{11}(x_1, y_1)$ ,  $M_{12}(x_1, y_2)$ ,  $M_{21}(x_2, y_1)$ ,  $M_{22}(x_2, y_2)$ .
- (3) Adjust the rotation of PTZ camera until the centre of PTZ camera image coincides with M11. Then read the current rotation angles in the horizontal direction and the vertical direction of PTZ camera  $(\alpha, \beta)_{11}$ , and record a group of data  $L[M_{11}(x_1, y_1)] = (\alpha, \beta)_{11}$ .
- (4) Repeat No. 3 operation for the rest intersections in ROI of wide angle image and take notes down all  $L[M(x, y)] = (\alpha, \beta)$ .



Figure 4. Dual cameras on the same vertical plane.

- (5) Look for the minimum rectangle  $M_{11}M_{12}M_{21}M_{22}$  encircling non-grid-intersection  $S(x, y)$  in ROI, and calculate the rotation angle of PTZ camera by means of bilinear interpolation formula.

$$(\alpha, \beta)_s = \frac{1}{(x_2 - x_1)(y_2 - y_1)} [L(M_{11})(x_2 - x)(y_2 - y) + L(M_{12})(x_2 - x)(y - y_1) + L(M_{21})(x - x_1)(y_2 - y) + L(M_{22})(x - x_1)(y - y_1)] \quad (1)$$

- (6) Combining data obtained in 3), 4), and 5) gives the look-up table for the PTZ camera rotating and aiming at arbitrary position of ROI in dome camera view.

When calibrating dual cameras in the way of Look-up Table, it is easy to operate with depth information of the object with regards to the two cameras changing a little. Therefore, the installation positions of the two cameras are as shown in Figure 4, that the lens of each camera approximately stays on the same vertical plane.

### 4.3. Face detection

CNN is multi-layer feed-forward architecture, which uses the supervised learning to extract invariant multi-stage features from image data. In CNN architecture, the individual neurons are tiled in the way that they respond to overlapping regions. Ideally, the “deep” representation would learn hierarchies of feature detectors and combine the top-bottom and bottom-up processing of an image. For instance, lower layers could support object detection. Conversely, information about objects in the higher layers could resolve the lower-level ambiguities. The structure of the CNN is illustrated in Figure 5, as follows:

Normally, a CNN is composed of several stages, as the example is showed in Figure 5, there are two stages. Each stage has a convolutional layer following with a non-linearity operation and a spatial feature pooling layer. The convolutional



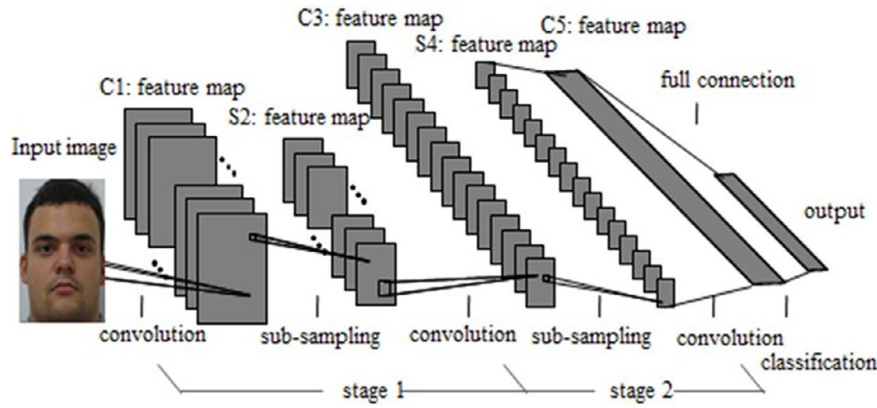


Figure 5. CNN structure used in face detection.

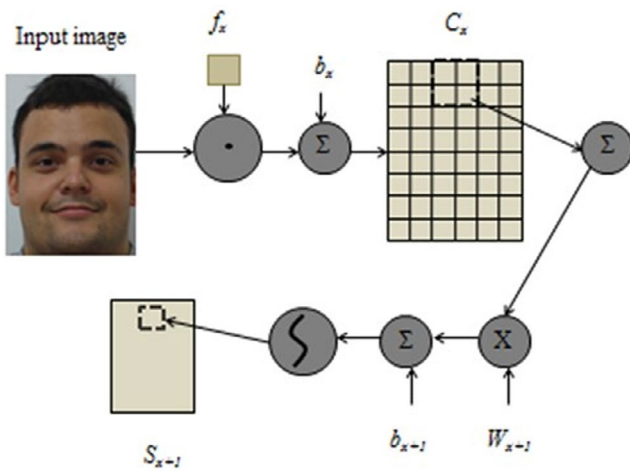


Figure 6. Illustration of convolution and subsampling in CNN.

layer consists of several trainable filter banks and additive bias. All the filters in filter banks can be trained. The filters with a specifically sized window are used to process the small local parts of the input image. The pooling layer lowers the spatial resolution by using a window of specific size, which leads to a strong robustness against geometric distortions. Normally the window size is smaller in the lower layers. Because in higher layers, in order to deal with the more complex part of the input image, a window with a bigger size is needed to get a lower resolution. Before the classification procedure, the features from all positions are combined and fully connected. The convolution and the subsampling process are illustrated in Figure 6.

In Figure 6,  $f_x$  stands for the trainable filters in the filter banks. First the input images are convoluted by using filters, and then the bias  $b_x$  are added, the convolutional layer  $C_x$  can be calculated. The subsampling procedure takes  $n \times n$  pixels, which are in the same neighborhood and merges them into one pixel, then adds weights  $W_{x+1}$  and bias  $b_{x+1}$ . The features could be projected to feature map  $S_{x+1}$  by using the sigmoid function.

After the two stages, the classification process can be accomplished by a classifier. Because of the extended hierarchical structure, CNN has the ability to learn not only the low-level features but also the mid-level features.

#### 4.4. Face orientation detection

Haar-like features are digital image features used in object recognition. They owe their name to their intuitive similarity with Haar wavelets and were used in the first real-time face

detector. Historically, working with only image intensities (i.e., the RGB pixel values at each and every pixel of image) made the task of feature calculation computationally expensive. A Haar-like feature considers adjacent rectangular regions at a specific location in a detection window, sums up the pixel intensities in each region and calculates the difference between these sums. This difference is then used to categorize subsections of an image. For example, let us say we have an image database with human faces. It is a common observation that among all faces the region of the eyes is darker than the region of the cheeks. Therefore a common Haar feature for face detection is a set of two adjacent rectangles that lie above the eye and the cheek region. The position of these rectangles is defined relative to a detection window that acts like a bounding box to the target object (the face in this case).

In the detection phase of the Viola–Jones object detection framework, a window of the target size is moved over the input image, and for each subsection of the image the Haar-like feature is calculated. This difference is then compared to a learned threshold that separates non-objects from objects. Because such a Haar-like feature is only a weak learner or classifier (its detection quality is slightly better than random guessing) a large number of Haar-like features are necessary to describe an object with sufficient accuracy. In the Viola–Jones object detection framework, the Haar-like features are therefore organized in something called a classifier cascade to form a strong learner or classifier.

The key advantage of a Haar-like feature over most other features is its calculation speed. Due to the use of integral images, a Haar-like feature of any size can be calculated in constant time.

A simple rectangular Haar-like feature can be defined as the difference of the sum of pixels of areas inside the rectangle, which can be at any position and scale within the original image. This modified feature set is called 2-rectangle feature. Viola and Jones also defined 3-rectangle features and 4-rectangle features. The values indicate certain characteristics of a particular area of the image. Each feature type can indicate the existence (or absence) of certain characteristics in the image, such as edges or changes in texture. For example, a 2-rectangle feature can indicate where the border lies between a dark region and a light region.

One of the contributions of Viola and Jones was to use summed area tables, which they called integral images. Integral images can be defined as two-dimensional lookup tables in the form of a matrix with the same size of the original image. Each element of the integral image contains the sum of all pixels located on the up-left region of the original image (in relation to the element's position). This allows computing the sum of rectangular areas in the image, at any position or scale, using only four lookups, see Figure 7.

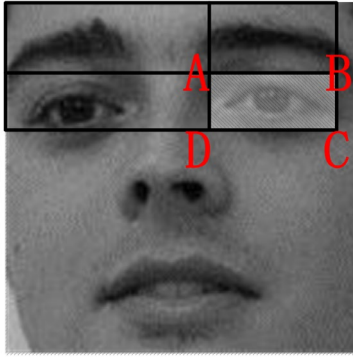


Figure 7. The sketch of computation on integral image.

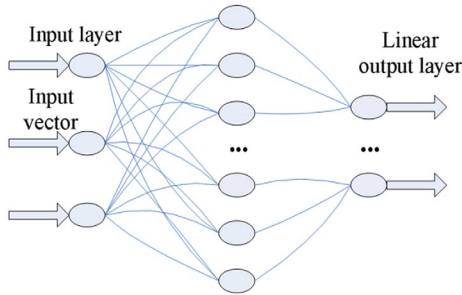


Figure 8. LVQ network model.

Equation 1 calculates the sum of the shaded rectangular area:

$$sum = I(C) + I(A) - I(B) - I(D) \quad (2)$$

where points A, B, C and D belong to the integral image  $I$ , as shown in Figure 7.

Each Haar-like feature may need more than four lookups, depending on how it was defined. Viola and Jones's 2-rectangle features need six lookups, 3-rectangle features need eight lookups, and 4-rectangle features need nine lookups. This was successful, as some of these features are able to describe the object in a better way. For example, a 2-rectangle tilted Haar-like feature can indicate the existence of an edge at  $45^\circ$ . Although the idea sounds mathematically sound, practical problems prevented the use of Haar-like features at any angle. In order to be fast, detection algorithms use low resolution images, causing rounding errors. For this reason, rotated Haar-like features are not commonly used.

LVQ network model is shown in Figure 8. A LVQ neuron network consists of three layers, i.e. input, competition and linear output. The network input layer is completely connected to the competition layer while the competition layer is partially connected to the linear output layer. Different connection exists between each output neuron group competition neuron group and the fixed value of them is 1.

## 5. Experiment

We utilized the real video surveillance system to verify the proposed gun-dome cooperative person-tracking system. The practical camera configuration refers to the left figure in Figure 9.

As to the above-mentioned cooperative calibration, we adopted the calibration toolbox on Matlab platform. The classical chessboard is utilized to get the internal parameter matrixes of the two component cameras, see Figure 9(a). The external

matrixes and the rotation and translation relation between the two cameras is calculated using the 3D calibration methodology, see Figure 9(b).

The performance of the proposed person-tracing architecture using gun-dome dual-camera system is superior to the single dome-based person tracing system. The prototype run effectively indoors and in the next stage we will deploy it outdoors and verify its effectiveness under partly occlusion.

The image data-set used for the face detection experiment was composed of two types of images, one image set including 100 facial images and the other including 100 non-face images. The image set was composed of the old, young, women and men face images and exists slightly illumination and orientation change. In the non-face image set, there are various kinds of furniture, buildings and natural scenes. 70 facial images and non-facial images are randomly selected from the two image sets for training respectively, and the left 30 facial images and 30 non-facial images were utilized to verify the performances of the CNN-based and LBP-SVM-based detection strategy. That is to say, the training samples and testing samples are manually categorized in 2 different classes which represent facial and non-facial image respectively.

The CNN that is used in present experiment consists of 7 layers, which is similar with the CNN illustrated in Figure 6. Layer C1 extracted 16 features from input images by using a  $5 \times 5$  kernel in each feature map. After subsampling, by using a  $4 \times 4$  kernel the layer C3 produces 16 features, layer C5 produces 120 features.

The training procedure of CNN is similar to the procedure of training BP neural network, which consists of two steps;

Step 1: Feed Forward

"Feed forward" describes the process of putting the input samples into the network, and then calculating the output.

$$O_p = F_n \left( \dots \left( F_2 \left( F_1 \left( X_p W^{(1)} \right) W^{(2)} \right) \dots \right) W^{(n)} \right) \quad (3)$$

$O_p$  is the practical output,  $W$  is the weight of the layer,  $F$  stands for the network layer.

Step 2: Back Propagation

The difference between practical output  $O_p$  and the ideal output  $Y_p$  will be calculated, and the weight matrix will be adjusted by minimizing the difference.

The adopted approach is based on CNN-based deep learning mechanism. The experiment used 8 kernels in layer C1, 16 kernels in layer C3, and 120 kernels in layer C5 to extract image features. A part of the outputs is showed in Figure 10. The CNN converged after the 4th iteration and the final accuracy rate at the 7th iteration over the test data-set is 100%.

To verify the effective of the proposed method for face orientation detection, we utilized it on real-time video stream and face image database respectively. For the simplicity, we mount the static analog camera on the tripod to test the effectiveness of the Haar-like feature on eye detection. Further study shows the proposed technique also has good performance on the IP camera and other types of surveillance cameras.

Figure 11 shows the result of eye detection in a range of head rotation. We can infer coarsely that the Haar-like feature is an advisable method to locate the eyes. The experiments shows that for the  $640 \times 480$  resolution video stream with the 25 fps, the consuming time on eye detection for one face image is less than 300 ms. Besides the above-mentioned photos acquired by the static camera, photos of 10 people (5 male and 5 female)

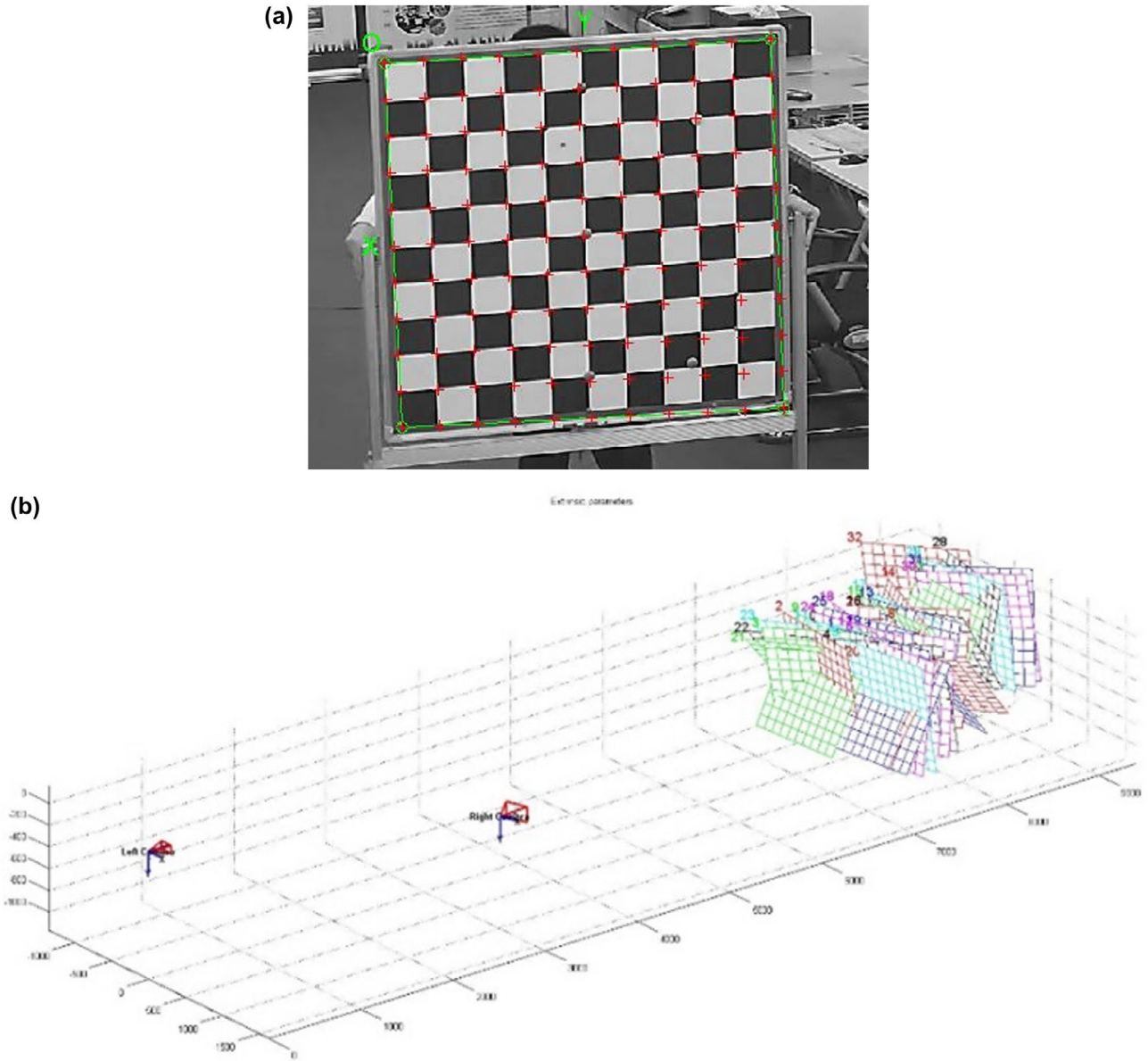


Figure 9. (a) The chessboard for calibration. (b) The correlation between camera coordinate and the chessboard coordinate.



Figure 10. The hierarchical structure of image features extracted by CNN.

were obtained from PIE Face Database of CMU from online website posted for scientific use. Each subject has five different head postures as “left”, “slight left”, “frontal”, “slight right” and “right”, see Figure 12.

As Figure 12 shows explicitly, when people face different orientations, the eye location and the distance between two eyes on the images will vary dramatically following the head rotating. Considering this fact, we adopt the statistic information about eye location as the input to the LVQ classifier. We denote the 5 orientations

as “1”, “2”, “3”, “4” and “5” respectively as the output of the LVQ classifier.

As Figure 13 shows, we divide the face image into 6\*8 sub-images and execute the edge detection by Canny operator, then accord the eye horizontal location denoted by the rectangle we calculate the sum of “1” pixels in the corresponding horizontal 8 sub-images. The sum is adopted as the feature feed to the LVQ classifier.

As the above mentioned, the total face images we got is 50 (10 persons, each person with 5 images). We randomly select





Figure 11. Eye location by using the haar-like feature in real-time video stream.

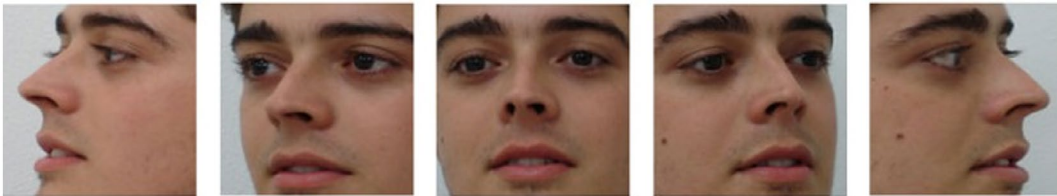


Figure 12. The illustration of 5 face orientations.



Figure 13. The illustration of eye location and the grid of sub-images.

30 images for training from the images. The left 20 images are used as the test samples to verify the performance of the proposed technique on face orientation detection. The classification experiment shows that for one face image, the consuming time is less than 600 ms. That is to say; the face orientation can be recognized in 0.6s.

## 6. Conclusions

In this paper, we propose a new gun-dome camera cooperative system, which adopts master-slave static wide angle - movable narrow angle dual cameras cooperative monitoring system.

There is one wide angle camera and one Pan Tilt Zoom (PTZ) dome machine. The wide angle camera is responsible for the target detection in wide field of view, and PTZ dome machine (also known as active camera) for focusing and amplifying and tracking continuously for the target of attention. It provides much better performance than the single dome camera. DPM and Look-up Table are used in this system. Furthermore, the deep-learning architecture and Haar-LQV are utilized to execute the face detection and orientation detection. The experiments prove the effectiveness and efficiency of the proposed scheme. In the future, the text based video search engine will be used for fast retrieval even under massive video data.



## Acknowledgement

Our research was supported by the Project of Shanghai Municipal Commission of Economy and Information (No.12GA-19), the standard revision project on public security named “Technical requirements for interested object detection and tracing using the collaborative multi-camera in surveillance video system,” (No. C14726), by the National Science and Technology Major Project under Grant 2013 ZX01033002-003, in part by the National Natural Science Foundation of China under Grant 61300202, 61300028, and in part by the Natural Science Foundation of Shanghai under Grant 13ZR1452900.

## Disclosure statement

No potential conflict of interest was reported by the authors.

## Notes on contributors



**Zhiguo Yan** received degree certificate of Doctor of Philosophy from Shanghai Jiaotong University in 2008, and accomplished the post-doctor research in Fudan University in 2013. He is a vice professor of the third research institute of Ministry of public security. His scholar interests focus on video intelligent analysis, big data techniques, IoT techniques on public security, etc.



**Zheng Xu** was born in Shanghai, China. He received Diploma and Ph.D. degrees from the School of Computing Engineering and Science, Shanghai University, Shanghai, in 2007 and 2012, respectively. He is currently working in the third research institute of ministry of public security and the postdoctoral in Tsinghua University, China. His current research interests include topic detection and tracking, semantic web and web mining. He has authored or co-authored more than 70 publications including IEEE Trans. on Fuzzy Systems, IEEE Trans. on Automation Science and Engineering, IEEE Trans. on Cloud Computing, IEEE Trans. on Emerging Topics in Computing, IEEE Trans. on Systems, Man, and Cybernetics-Part C, etc.



**Jie Dai** is an assistant professor at the Department of Internet of Things, The Third Research Institute of the Ministry of Public Security. He received his Ph.D. in Computer Science and Engineering from Hong Kong University of Science and Technology. His research interests include big data analysis and video analysis in large-scale systems.

## ORCID

Zheng Xu  <http://orcid.org/0000-0002-8362-5991>

## References

- Beriault, S. (2008). *Multi-camera system design, calibration and 3D reconstruction for markerless motioncapture*. (Thesis). School of Information Technology and Engineering, Engineering University of Ottawa, p. 146.
- Cho, H., Rybski, P.E., Bar-Hillel, A., & Zhang, W. (2012). Real-time pedestrian detection with deformable part models. In *Intelligent Vehicles Symposium IEEE: Alcalá de Henares, Spain*, pp. 1035–1042.
- Dong, R., Li, B., & Chen, Q.-M. (2009). An automatic calibration method for PTZ camera in expressway monitoring system. In *WRI World Congress on Computer Science and Information Engineering*, 636–640.
- Felzenszwalb, P., McAllester, D., & Ramanan, D. (2008). A Discriminatively Trained, Multiscale, Deformable Part Model. In *Computer Vision and Pattern Recognition* (pp. 1–8). Anchorage, AK: IEEE.
- Hao, Z., Zhang, X., & Yu, P. (2010). *Video Object Tracing Based on Particle Filter with Anti Colony Optimization*. In *The 2nd IEEE International Conference on Advanced Computer Control*, Shenyang, China. pp. 232–236.
- Kim, J.-M., Kim, K.-H., & Song, M.-K., 2009. *Real time face detection and recognition using rectangular feature based classifier and modified matching algorithm*. In *2009 Fifth International Conference on Natural Computation*, Korea, pp. 171–175.
- Li, H., & Shen, C., 2006. An LMI approach for reliable PTZ camera self-calibration. In *International Conference on Advanced Video and Signal Based Surveillance (AVSS'06)*, IEEE.
- Xie, D., Dang, L., & Tong, R., 2012. *Video based head detection and tracking surveillance system*. In *International Conference on Fuzzy Systems and Knowledge Discovery (FSKD) IEEE, Sichuan*, pp. 2832 – 2836.
- Xu, Z., & Chen, H. (2015a). The semantic analysis of knowledge map for the traffic violations from the surveillance video big data. *Computer Systems Science & Engineering*, 30(5), 403–410.
- Xu, Z., Liu, Y., Mei, L., Hu, C., & Chen, L. (2015b). Semantic based representing and organizing surveillance big data using video structural description technology. *Journal of Systems and Software*, 102, 217–225.
- Xu, Z., Mei, L., Hu, C., & Liu, Y. (2016a). The big data analytics and applications of the surveillance system using video structured description technology. *Cluster Computing*, 19, 1283–1292.
- Xu, Z., Mei, L., Liu, Y., Hu, C., & Chen, L. (2016b). Semantic enhanced cloud environment for surveillance data management using video structural description. *Computing*, 98, 35–54.
- Xu, Z., Hu, C., & Mei, L. (2016c). Video structured description technology based intelligence analysis of surveillance videos for public security applications. *Multimedia Tools and Applications*, 75, 12155–12172.