Real Time Door Access Event Detection and Notification in a Reactive Smart Surveillance System

Gaetano Di Caterina, Nurulfajar Abd Manap, Masrullizam Mat Ibrahim,
John J. Soraghan

Department of Electronic and Electrical Engineering, University of Strathclyde,
Royal College Building, 204 George Street, Glasgow, G1 1XW, UK
gaetano.di-caterina@strath.ac.uk

Abstract. The effectiveness of modern video surveillance systems critically depends on camera image quality and human operators' reactivity. In this paper we present a door access event detection application in the context of a reactive smart surveillance system, which automatically notifies in real time the occurrence of events to registered users, through SMS alerts. The system utilizes two fixed IP cameras and a high resolution PTZ camera to acquire high quality images of the face of people entering the room. System users can access a web-based interface to review the event details, along with a short video clip and the high quality face images acquired. Experimental results demonstrate that the final system allows the PTZ camera to automatically acquire high-resolution images of faces and deliver them to system operators in real time.

1 Introduction

CCTV is not always as effective as expected, due to two important issues, namely (i) image quality and (ii) reactivity of the surveillance personnel in spotting events of interest. To address these issues, digital video surveillance systems are beginning to incorporate megapixel IP cameras, which can deliver high quality images over IP networks, at high frame rate. Secondly, smart technologies can be used to analyze the video feeds and detect events of interest in real time for effective use. In smart surveillance systems [1], video analytics, which is the semantic analysis of video data through signal and image processing techniques, is used to extract and process only the relevant information, to reduce both processing time and storage space.

The main contribution of this paper is the incorporation of a door access event detection application in the context of the reactive smart surveillance system described in [2]. In particular the proposed system can automatically detect and record faces of people entering a room, and notify the door access events in real time to registered users, through SMS alerts. Two low resolution IP cameras are used to obtain the 3D location of the object of interest, which is the face of people entering the room, with stereo matching techniques [3]. Such positional information is passed to a high resolution pan-tilt-zoom (PTZ) camera to locate the detected face and acquire high quality images of it. The presented system builds on the work in [4]. However, the system in [4] only focused on static objects, while the surveillance system proposed in this paper is integrated within a door access monitoring framework, wherein it can detect and record

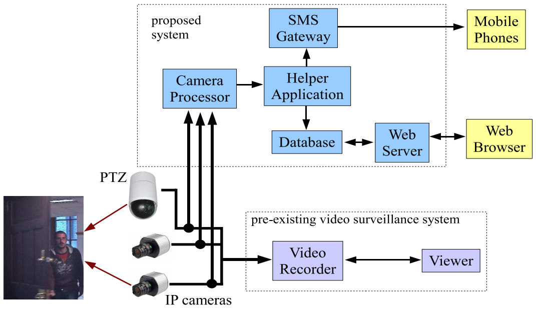


Fig. 1. System design with two IP cameras and a PTZ camera.

moving objects, such as the faces of people entering the room. The remainder of the paper is organized as follows. Section 2 gives a brief overview of the system architecture. Section 3 provides a detailed description of the system camera processor. Section 4 contains experimental results and discussion, while section 5 concludes the paper.

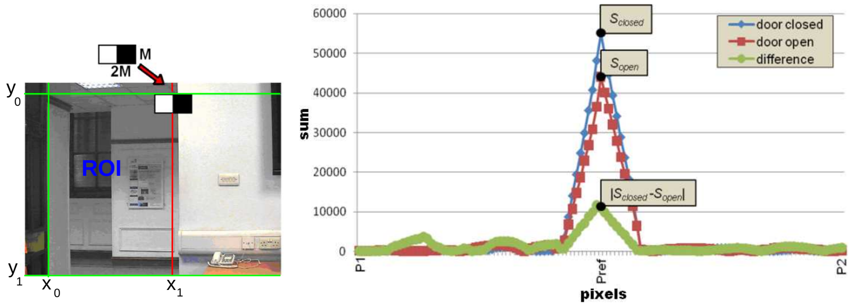
2 System architecture

A block diagram of the overall reactive smart surveillance system is depicted in Fig. 1. The system hardware includes two 1.3 megapixel Arecont AV1300 fixed IP cameras, and a 5 megapixel ACTi IP Speed Dome (CAM-6510) PTZ camera. The system software components are: one camera processor, which analyzes the input video feeds; a web server and associated database to store details of the detected events; a helper application which saves event data received from the camera processor into the database and sends SMS alerts to registered users. The camera processor is implemented in Matlab, Java and C, and it includes the video analytics algorithms, the PTZ controller and the event notification block. The two IP cameras are set up in a stereo configuration and have the door in their field of view. When the door opens, the IP cameras acquire real time images from two different angles. Such images are combined to produce stereo vision and compute the 3D location of the object of interest, i.e. the face of the person entering. This information is fed to the PTZ controller, which pans and tilts the PTZ camera to point at the targeted face and acquire a high resolution image of it. The door access event is also notified in real time to registered users, through SMS alerts.

3 Camera processor description

3.1 Door open detection

The main objective of this block is to detect in each new frame whether the door is open or closed. Door open detection is performed only on the left image, for simplicity. Since the two IP cameras are fixed, it is reasonable to select a region of interest (ROI) for the door in the $W \times H$ image, either manually or automatically [5], as shown in Fig. 2(a), with x_0, x_1, y_0 and y_1 being the horizontal and vertical coordinates of the ROI. The



(a) (b)

Fig. 2. Door open detection. (a) $2M \times M$ binary mask applied to the door image; (b) Behaviour of the sum S_i in both 'door open' and 'door closed' images.

vertical side of the ROI where the hinges of the door are, is identified as 'hinge side', while the other vertical side is identified as 'free side'. In order to detect whether the door is open in the i^{th} frame, a $2M \times M$ binary mask as in Fig. 2(a) is overlapped across the 'free side' at the top, in position $\mathbf{P}_{ref} = (x_1 - M, y_0)$, so that no object can ever occlude this part of the ROI. In usual video surveillance setups, cameras are mounted from the ceiling or at the very top of side walls, therefore the line of sight between camera and top edge of the door is never occluded. The pixel values in the binary mask are multiplied with the corresponding pixel values in the i^{th} frame and summed together to obtain the sum S_i at position $\mathbf{P}_{ref}$. As an experiment, if the binary mask scans the 'door closed' and 'door open' images horizontally, with its position going from $\mathbf{P}_1 = (x_1 - 3M, y_0)$ to $\mathbf{P}_2 = (x_1 + 3M, y_0)$, the graph in Fig. 2(b) is obtained. It is possible to see that in position $\mathbf{P}_{ref}$ the sum S_i can assume two very different values S_{open} and S_{closed}, when the door is respectively open and closed. The only assumption here is that the door, the wall beside it and the background behind it do not all have the same colour. This suggests that a threshold χ can be set as $\chi = |S_{open} - S_{closed}|/2$. For the i^{th} frame, S_i is computed and if $|S_i - S_{closed}| > \chi$, then the door is considered to be open and the face detection algorithm is run. The presented door open detector is simple and fast and it can be seen as an improved motion detector that works on the underlying image structure: in a conventional motion detector, the pixel-wise difference between frames is thresholded to detect motion; in the proposed detector, the strength of the vertical edge on the door 'free side' is analyzed instead. Therefore, while a conventional motion detector could also be triggered by shadows and light changes, the presented door open detector is triggered only when the door is actually open, i.e. the strength of the vertical edge on its 'free side' varies.

3.2 Face detection

There are four stages in the face detection step: skin colour segmentation, morphological processing, bounding rectangle forming and SVM classification. The obvious advantages of skin colour segmentation are fast processing and high robustness to geometric variation of head pose and orientation. For this purpose three colour spaces have been employed: RGB, YCbCr and HSV. These three colour spaces are widely used in skin detection research [6–9]. RGB is the most used one, although it is not very robust to light changes. Therefore Kovac *et al.* [7] used gray world method as an adaptation technique, to correct the images before applying skin detection. To adapt to light changes, Pai *et al.* [8] modulated the range of YCbCr skin colour distribution. The last colour space, i.e. HSV, represents colours in terms of depth, purity and brightness [6,9]. From these three colour spaces, a combination rule for segmentation is formulated as in (1), to overcome sensitivity to illumination changes, ethnicity skin colour and different characteristics of cameras.

$$\begin{aligned}
&\text{if } (r > 95 \wedge g > 40 \wedge b > 20)\\
&\quad \wedge ((\max(r,g,b) - \min(r,g,b)) > 15)\\
&\quad \wedge (|r-g| > 15) \wedge (r > g) \wedge (r > b)\\
&\quad \wedge (140 < c_b < 195) \wedge (140 < c_r < 165)\\
&\quad \wedge (0.01 < hue < 0.1)\\
&\text{then } \textit{(selected pixel is skin)}
\end{aligned} \tag{1}$$

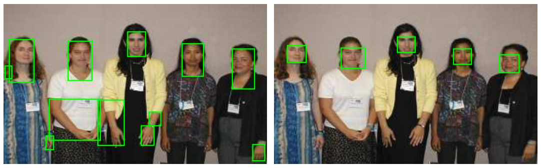
To obtain well segmented skin regions, mathematical morphology is used to remove noise and fill small holes. Bounding rectangles are formed by using the connected components labeling operator. Each bounding rectangle is then examined in terms of size and pattern. The pattern shape describes whether the rectangle bounds a face or a non-face object, and it is measured by the width-to-height ratio of the rectangle defined as:

$$0.83 < \frac{width}{height} < 1.27 \tag{2}$$

The range values in (2) are obtained from experiments carried out on 98 images containing 561 faces. Fig. 3(a) shows example of experimental results after skin colour segmentation and rectangle bounding formation. In these images, the rectangles bound all the regions segmented as skin. Rectangles that are too small or do not comply with (2) are discarded, as in Fig. 3(b). The remaining bounding rectangles are then classified as whether containing face or non face by using SVM on horizontal projection features. The horizontal projection of a face has a distinctive pattern, which is used as features for SVM training and classification. Fig. 4 shows three different poses of face, with horizontal projection profiles of eyes, nose and mouth. The values of peak and valley projected by the horizontal profile are used as features to differentiate between face and non-face objects. For this purpose, the image regions included in the remaining bounding rectangles are converted to gray scale. However, due to noise, such regions project an indistinctive horizontal graph projection, from which it is difficult to extract features. Therefore a Gaussian filter is employed to smoothen such face candidate regions. These smoothened regions are finally classified using SVM. Face regions in output from the SVM classifier in the left image are then processed in the stereo matching step, to find their corresponding regions in the right image.

3.3 Stereo matching and 3D location estimation

Stereo matching determines which parts of the left and right images correspond to the same scene element. The central block from the detected human face in the left image



(a) (b)

Fig. 3. Experimental results after skin colour segmentation and rectangle bounding formation.

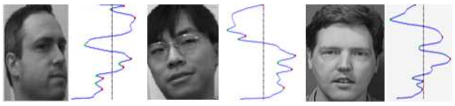


Fig. 4. The horizontal projection profile of three faces.

is taken as a reference and compared with blocks in a search area in the right image. The block size is constrained to $\psi \times \psi$ pixels, while the size of the search area is of $\xi \times \xi$ pixels. The actual values of ψ and ξ depend on the application and also on the stereo camera setup. In the proposed system these values are $\psi = 32$ and $\xi = 128$. The matching between blocks in the left and right cameras is determined by the value of a cost function. Here, any matching measure could be used; however for low computation, we use the Sum of Absolute Differences (SAD). Minimizing the SAD measure gives the position in the right image of the best match for the reference block selected in the left image. To calculate the accurate 3D location of the detected human face, basic geometry rules are used. The simplest geometry of stereo video system consists of two parallel cameras with horizontal displacement, i.e. along the X axes, as shown in Fig. 5. Such geometry is derived from the pinhole camera model [10] and the same horizontal line is referred to as epipolar line. The symbol f is the focal length of the camera lens and B is the baseline distance, i.e. the distance between the two camera optical centres. If $\mathbf{O}_L = (U_L, V_L)$ and $\mathbf{O}_R = (U_R, V_R)$ are the projections in the left and right images, relative to the respective camera centre points, of the 3D point $\mathbf{P}_L$, as illustrated in Fig. 5, it holds $V_L = V_R$ and the disparity of the stereo images is obtained as difference between U_L and U_R:

$$d = U_L - U_R = \left(f\frac{x_L}{z_P} - f\frac{x_R}{z_P} \right) = \left(f\frac{x_L}{z_P} - f\frac{x_L - B}{z_P} \right) \tag{3}$$

The location of correct projections of the same point $\mathbf{P}$ on the two image planes can determine the exact depth of $\mathbf{P}$ in the real world. From (3), the depth z_P of the point $\mathbf{P}$ is computed as $z_P = (fB)/d$. Therefore, the equations used to calculate the exact location $\mathbf{P} = (x_P, y_P, z_P)$ of the target object are:

$$x_P = \frac{Bx_L}{d}, \qquad y_P = \frac{By_L}{d}, \qquad z_P = \frac{Bf}{d} \tag{4}$$

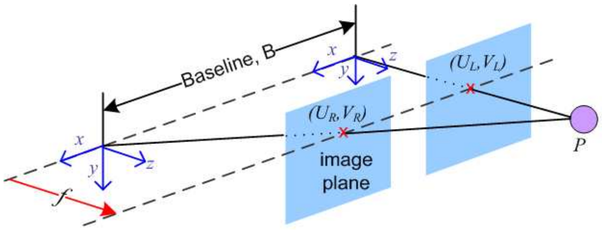


Fig. 5. Stereo camera configuration.

3.4 PTZ controller

The PTZ controller module deals with the PTZ hardware, firmware and communication protocols. First, it applies a homogeneous transformation to compute the 3D location $\mathbf{P}_{PTZ} = (x_{PTZ}, y_{PTZ}, z_{PTZ})$ of the target with respect to the PTZ. If $\mathbf{T}$ is a transformation matrix that transforms from the stereo cameras coordinate frame to the PTZ coordinate frame, the location $\mathbf{P}_{PTZ}$ is computed as:

$$[x_{PTZ}, y_{PTZ}, z_{PTZ}, 1]^T = \mathbf{T}\,[x_P, y_P, z_P, 1]^T \tag{5}$$

The PTZ controller converts the target location $\mathbf{P}_{PTZ}$ into pan and tilt angles, and zoom factor for the PTZ. These values are incorporated into commands for the PTZ, in the form of standard HTTP requests, over the network. The panning angle θ and the tilting angle β are calculated as:

$$\theta = \tan^{-1}\left(\frac{z_{PTZ}}{D - x_{PTZ}}\right) \tag{6}$$

$$\beta = \tan^{-1}\left(\frac{y_{PTZ}}{\sqrt{(D - x_{PTZ})^2 + z_{PTZ}^2}}\right) \tag{7}$$

where D is the distance between IP cameras and PTZ along the X axes. The zoom ratio instead is proportional to the Euclidean distance between PTZ camera and target object.

3.5 Event notification

When the door is detected as open as described in section 3.1, a timer is started and after 10s a door access event is triggered. At this point a low frame rate $(2-5\text{fps})$ video clip of the past 10s is created and asynchronously sent to the helper application, along with event details, such as time, date and camera ID. The helper application saves the event data in the web server database and issues an SMS alert to a list of pre-registered users, who can access the remote interface, to review event details and short video clip in real time, along with the high resolution face images recorded by the PTZ camera. The time delay before triggering a door access event is needed to make sure that the short video clip includes also images of the actual person entering the room. Within such interval, no other events are triggered. This is to prevent events from being triggered at every frame. However, if the door stays open for more than 10s, the timer is started again and a new event is triggered when the timer expires again.

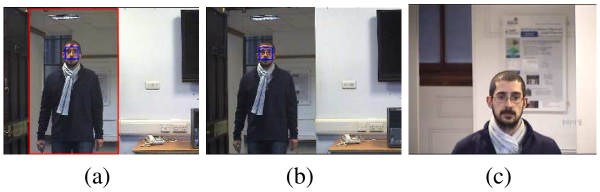

(a) (b) (c)

Fig. 6. Visual results. (a) left image; (b) right image; (c) high resolution image from the PTZ.

Axes	X	Y	Z
μ	0.047m	0.099m	0.357m
σ	0.027m	0.011m	0.077m

Table 1. Mean and standard deviation of absolute differences.

Operation	μ	σ
Acquisition	0.090s	0.004s
Face detection	0.032s	0.002s
Stereo matching	0.028s	0.001s
Location estimation	0.001s	0.000s

Table 2. Average execution times.

4 Results and discussion

Fig. 6 shows results of the face detection and high resolution face image acquisition. Fig. 6(a) and (b) are the left and right camera views and they are in the same epipolar line. The searching area is minimized to the door mask region only, instead of all the pixel images. With this approach, the execution of stereo matching and face detection is faster. The distance between IP cameras and the target in Fig. 6 is 4m. The face detection algorithm was tested with the CMU face colour images database [11], which contains a variety of faces in normal room lighting conditions. 346 face images with a variety of skin colour tones and different facial poses were used. The face detection described in this paper correctly detected human faces in 327 images (94.5%), with 19 images (5.5%) erroneously detected. The main cause of the errors was due to pieces of clothing classified as skin.

The face detection result is processed in the block matching and 3D location estimation steps, to obtain the depth and location of the targeted object. With this information, the coordinates of the object are calculated and transmitted to PTZ camera controller. The coordinates are converted into pan and tilt angles for the PTZ. The PTZ camera captures the targeted object as shown in Fig. 6(c), where the distance between object and PTZ camera is calculated as 8.13m. The object detected with the PTZ can be tracked and images of it are recorded automatically. The system has been developed and tested using different test vectors, by placing the cameras at different locations with respect to the PTZ, and with different people as target. The PTZ response upon changes of the coordinates has been found to be quick.

For the location estimation test, the system is fed with the 22 sets of stereo images, to evaluate the accuracy of the target location estimated by the proposed system, with respect to the exact target location in the 3D space. The error between each set of estimated and exact locations is computed as Euclidean distance. Table 1 shows means

μ and standard deviations σ of the absolute differences between exact and estimated values, for each coordinate axis. The error in X and Y coordinates are very small, while the error in Z coordinate is slightly higher.

The mean and standard deviation profile of the recorded execution times are presented in Table 2. The results show that face detection, stereo matching and location estimation steps accounts for less than 50% of the total execution time of 133ms. The high image acquisition time is due to the transmission of both left and right images over the network, from the IP cameras. The average frame rate is about 8fps. It is expected that an implementation on a dedicated DSP board would significantly speed up the total execution time.

5 Conclusion

A fully automated reactive smart surveillance system using stereo images has been designed and developed. It automatically detects door access events and uses multiple cameras to localize and zoom in on the faces of people entering the room, to acquire high resolution images of them. System features include door detection, face detection, high quality face image acquisition and real time notification to registered users. The overall system makes extensive use of IP technologies, to ensure communication among components and remote availability of the system resources, such as IP cameras, event database and user front-end. Despite its simplicity, the proposed system performs well and it is suitable for real time execution. As future work, the presented video analytics algorithms will be ported to a DSP board for fast 'in-camera' processing.

References

1. Valera, A., Velastin, S.A.: Intelligent distributed surveillance systems: a review. IEE Proceedings - Vision, Image and Signal Processing **152** (2005) 192–204
2. Di Caterina, G., Soraghan, J.J.: An abandoned and removed object detection algorithm in a reactive smart surveillance system. In: DSP2011. (2011)
3. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. Int. Journal of Computer Vision **47** (2002) 7–42
4. Manap, N.A., Di Caterina, G., Soraghan, J.J., Sidharth, V., Yao, H.: Smart surveillance system based on stereo matching algorithms with IP and PTZ cameras. In: 3DTV-Con2010. (2010) 1–4
5. Yang, X., Tian, Y.: Robust door detection in unfamiliar environments by combining edge and corner features. In: IEEE CVPR Workshops. (2010) 57–64
6. Chaves-Gonzalez, J.M., Vega-Rodriguez, M.A., Gomez-Pulido, J.A., Sanchez-Perez, J.M.: Detecting skin in face recognition systems: a colour spaces study. Digital Signal Processing **20** (2010) 806–823
7. Kovac, J., Peer, P., Solina, F.: Human skin color clustering for face detection. In: IEUROCON 2003 - Computer as a Tool. (2003) 144–148
8. Pai, Y.T., Ruan, S.J., Shie, M.C., Liu, Y.C.: A simple and accurate color face detection algorithm in complex background. In: IEEE ICME. (2006) 1545 –1548
9. Kakumanu, P., Makrogiannis, S., Bourbakis, N.: A survey of skin-color modeling and detection methods. Pattern Recognition **40** (2007) 1106–1122
10. Bovik, A.: Handbook of image and video processing. 2 edn. Elsevier, Academic Press (2005)
11. CMU: Image data base: face. (http://vasc.ri.cmu.edu/ idb/html/face/frontal_images/)