SMART SURVEILLANCE SYSTEM BASED ON STEREO MATCHING ALGORITHMS WITH IP AND PTZ CAMERAS

Nurulfajar Abd Manap, Gaetano Di Caterina, John Soraghan, Vijay Sidharth, Hui Yao

CeSIP, Electronic and Electrical Engineering, University of Strathclyde, UK

ABSTRACT

In this paper, we describe a system for smart surveillance using stereo images with applications to advanced video surveillance systems. The system utilizes two smart IP cameras to obtain the position and location of objects. In this case, the object target is human face. The position and location of the object are automatically extracted from two IP cameras and subsequently transmitted to an ACTi Pan-Tilt-Zoom (PTZ) camera, which then points and zooms to the exact position in space. This work involves video analytics for estimating the location of the object in a 3D environment and transmitting its positional coordinates to the PTZ camera. The research consists of algorithms development in surveillance system including face detection, block matching, location estimation and implementation with ACTi SDK tool. The final system allows the PTZ camera to track the objects and acquires images in high-resolution quality.

Index Terms — IP cameras, stereo vision, image matching, object detection, intelligent systems, surveillance

1. INTRODUCTION

Intelligent video surveillance systems have become more important and widely used in the last few years due to the increasing demand for safety and security in many public environments including transport applications, crowded public places and surveillance of human activities [1]. The growing need of security has emphasised the use of smart technology in surveillance systems. There are different types of technologies used for the purpose of tracking and location estimation, ranging from satellite imaging to CCTV. While satellite imaging is mostly employed by the military industry and it is expensive and used for wider scale, CCTV covers limited areas and is relatively cheaper, but requires constant monitoring by additional personnel for the purpose of detecting suspicious activity.

The amount of video surveillance data that is captured daily is growing exponentially. Scalability and usability become very critical to handle this huge amount of information. The availability and cost of high resolution surveillance cameras, along with the growing need for remote controlled security, have been a major driving force in this field. A growing amount of information increases the demand on processing and tagging this information for subsequent rapid retrieval. To satisfy this demand, many researches are working to find improvements and better solutions in video analytics, which is the semantic analysis of video data, to reduce running time and total cost of surveillance systems. Video analytics for surveillance basically involves detecting and recognizing objects in an automatic, efficient fashion. In smart surveillance systems, only important data are extracted from the available video feeds and passed on for further processing. This can save both time and storage space, and it makes video data retrieval faster, as significantly less data have to be scanned through with the use of smart tags.

978-1-4244-6379-4/10/$26.00 ©2010 IEEE

Stereo matching continues to be an active research area [2,3,4]. The main aim of stereo matching is to determine disparities that indicate the difference in locating corresponding pixels. Disparity maps allow estimating 3D structure of the scene and the geometry of the cameras in space. Many techniques have been proposed in order to determine the homologous points of the stereo pair. Scharstein and Szelinski [5] provided a valuable taxonomy and evaluation of dense stereo matching algorithms for rectified image pairs. Besides its application for 3D depth map, the stereo matching algorithm has been used as one of the key element in our system.

The main contribution of this paper is the presentation of a smart surveillance system for human face tracking. Multiple IP cameras have been used to obtain the 3D location of the object. Its positional information is passed to the PTZ camera to locate the targeted object. Stereo matching algorithms, normally used to obtain the depth map for 3D video and free-viewpoint video, have been exploited. They include adaptive illumination compensation, skin color segmentation, morphological processing and region analysis.

2. SYSTEM OVERVIEW

The layout of the proposed system is shown in Figure 1, with two IP cameras and one PTZ camera, which can pan of 360°, tilt of 90° and zoom. Even if the range of view of the PTZ is quite broad, it still requires a human operator to control it, sending commands through a web-based user interface.

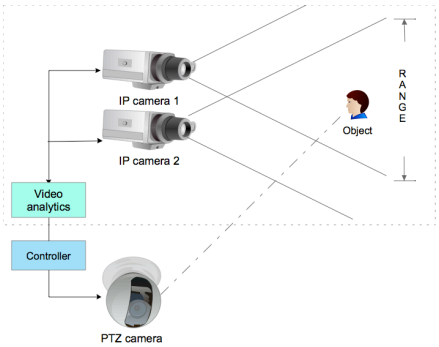


Figure 1. System design with two IP cameras and a PTZ camera

One of the main purposes of this research is to find an efficient approach to enable the PTZ camera to automatically detect objects and track them. Therefore, the two fixed IP cameras are used to acquire real-time images, which are processed by the video analytics algorithms to estimate the location of the object of interest. The IP cameras capture the same scene from two different angles; therefore the two acquired video streams can be combined to produce stereo video. The 3D coordinates are computed by the stereo matching and location estimation algorithms, and then they are fed to the PTZ controller to point at the object location.

In this implementation, the IP cameras used are two Arecont AV3100 3.1 mega pixels, which can capture frames of 2048x1536 pixels at 15 fps. Meanwhile, the PTZ camera is 5-mega pixels ACTi IP Speed Dome (CAM-6510). A very important feature is its capability of panning and tilting at 400° per second, which makes it very suitable for tracking. It can respond very quickly to any changes, with several movements and wide area of coverage. This PTZ camera supports both PAL and NTSC standard, being capable of transmitting at a rate of 25 fps and 30 fps respectively.

The system can be divided into two main subsystems, which are the video analytics block and the controller block. In the video analytics subsystem, the algorithms employed consist of face detection, stereo matching and location estimation. The face detection algorithms include adaptive illumination compensation, skin color classifier, morphological processing and connected region analysis. The second subsystem acts as the controller for the PTZ, dealing with its hardware, firmware and protocols. In this paper, the RGB color space is chosen. A decision rule based on the RGB color space is adopted, to explicitly define skin color cluster boundaries and discriminate between skin/non-skin pixels, as suggested in [6].

The main feature of the system is that it calculates the exact location of a detected object, which in this case is the human face. In this scenario, the real world coordinate system coincides with the left camera's coordinate system. The video analytics block can be simplified as in Figure 2. The system attempts to detect a human face in both images acquired by the two IP cameras, and then determines the central blocks of the human face in both images. The central block C_1 from the first image is selected as reference and the algorithm searches for the best matching block around C_2 in the second image. The projection of the same point on the two image planes of the two pixels in block C_1 and C_2 is used to estimate the location of the point P. The next section discusses the video analytics algorithms in more details.

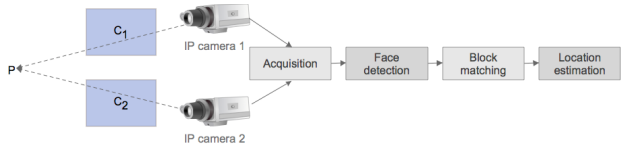


Figure 2. Video analytic subsystem

3. VIDEO ANALYTICS ALGORITHMS

The first part of the proposed system is the video analytics block, which consists of face detection, stereo matching and location estimation algorithms.

3.1 Face Detection

The main function of this sub-block illustrated in Figure 2 is to segment the face region out from the input image. The steps in the face detection algorithms are adaptive illumination compensation, skin color segmentation, morphological processing and connected component analysis. For this research, the skin color is used as main distinguishing feature in face detection. The obvious advantages of the skin color segmentation are fast processing and high robustness to geometric variation of head pose and orientation. RGB color space is chosen and explicit skin color cluster boundaries are defined, to discriminate between skin and non-skin pixels.

Due to the different lighting conditions in different environments, the appearance of human skin color can change, obviously affecting the skin region segmentation result. Therefore, the "Gray World Assumption" method [7] is used to perform adaptive illumination compensation. This method assumes that the average value of the RGB components in an image should average out to a common gray scale value. Each color component is scaled according to the amount of its deviation from this gray value.

A skin tone model is therefore defined. The classification of skin tone is taken from research work [8,9]. The skin color segmentation rejects non-skin regions and retains skin regions. Morphological processing is used to reduce noise. It uses the opening operation to reject tiny object and closing operation to fill tiny holes. The connected region analysis is then performed to reject non-face regions.

In this research, two geometry features of human face are examined to reject non-face regions: region ratio and roundness. The height to width ratio of human face should satisfy some specific relationship, in this case around 1. Therefore, if the height to width ratio of a connected region R satisfies $0.8 \le R \le 2.2$, the region will be a candidate region for the next step.

The shape of human face can be seen as an ellipse from different angles. The roundness of a connected region can be used to reject non-face regions, according to (3.1):

$$C_i = \frac{A_i}{P_i^2} \tag{3.1}$$

where A_i is the area of the i^{th} connected region, P_i is the perimeter of the i^{th} region. If $C_i > \tau$, with $\tau = 0.05$, the connected region is retained for the next step.

Besides the geometry features, also holes are a useful feature to classify a skin region as a human face [6]. The idea behind this is to find a region containing at least two holes, which correspond to the two eyes. The number of holes in a connected region can be calculated by computing the Euler number of the region. It is defined as:

$$E = C - H \tag{3.2}$$

where E is the Euler number, C is the number of connected components and H is the total number of holes in them. In the connected region analysis, C is set to 1 because only one connected region is analyzed at a time. If the Euler number of a connected region is $E < 0$, the region is rejected.

(a) (b)

Figure 3. Morphological processing and connected region analysis. (a) Image after morphological processing. (b) Region of interest

As illustrated in Figure 3, images may contain non-face objects that have similar colour as the human skin, such as the cupboard on the left-hand side of Figure 3(a). All the white pixels regions shown in Figure 3(a) are considered human skin regions. It is obvious that not all of them contain holes, except region 5. Besides, the height to weight ratio and roundness of region 2 and 3 exceed the given thresholds. Therefore, only region 5 is classified as a face region. After the analysis, the non-face regions are rejected and the coordinates of the face regions, in both left and right images, are reserved for stereo matching algorithms as shown in Figure 3(b).

3.2 Stereo Matching Algorithms

The stereo matching is the fundamental step in determining which parts of two images are projections of the same scene element. The main aim of stereo matching algorithms is to find homologous points in the stereo pair [10]. In the first stage of the research, a block matching algorithm is used. These techniques are widely used in motion estimation. The idea behind motion estimation is that the current frame is divided into several macro blocks. Then a block-matching algorithm is used to compare macro blocks in the current frame with the corresponding blocks and its adjacent neighbors in the previous frame, to create a motion vector, which describes the movement of a macro block from one location in the previous frame to the current location. The basic idea of the block matching algorithms is used for stereo matching between the pair of images, to estimate the depth and location of the targeted objects.

In this research, a one-step search [10] is selected for faster execution and low complexity. A central block of human face in the left image is taken as a reference and compared with another block in the target image, which is the right image. The process of searching for similar matching block is constrained to 16x16 pixels for the size of macro block match, and the size of 7 pixels for the search area. The matching between reference block and the target block is determined by the value of a cost function. Here, any matching measure could be used; however, again for low computation, we use the Sum of Absolute Differences (SAD), which is given by in the following equation:

$$SAD = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left|C_{ij} - R_{ij}\right| \tag{3.3}$$

where N is the size of the macro block, C_{ij} and R_{ij} are the pixels of the target and reference macro block respectively.

3.3 Location Estimation Algorithms

In order to calculate the accurate 3D location of the detected human face, basic geometry rules are used. The projection of a 3D physical point onto the two image planes requires finding the exact location of the object [10]. The simplest geometry of stereo system consists of two parallel IP cameras with horizontal displacement as shown in Figure 4. The stereo configuration is derived from the pinhole camera model [11].

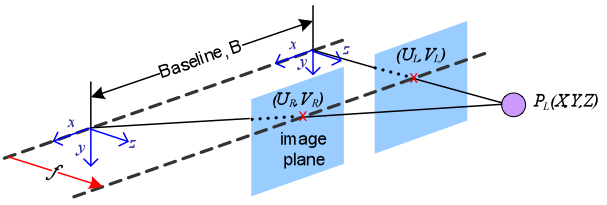


Figure 4. Stereo camera configuration

Referring to Figure 4, O_L is the reference camera centre point (or the left camera), while O_R is the target camera centre point. The implementation of this system is based on parallel cameras, which are shifted along the same horizontal line or x-coordinate, known as the epipolar line. Therefore, v_L=v_R. The symbol f is the focal length of cameras' lens (the distance from camera centre point to the image plane) and B is the baseline distance (distance between two optical centers, O_L and O_R). The points of the images can be described as the following:

$$(u_L, v_L) = \left(f\frac{x}{z}, f\frac{y}{z}\right) \tag{3.4}$$

$$(u_R, v_R) = \left(f\frac{x-B}{z}, f\frac{y}{z}\right) \tag{3.5}$$

The disparity of the stereo images is obtained as difference between the two corresponding points, U_L and U_R:

$$disparity,\ d = u_L - u_R = \left(f\frac{x}{z} - f\frac{x-B}{z}\right) \tag{3.6}$$

The location of correct projections of the same point P_L on the two image planes can determine the exact depth of P_L in the real world. From equation (3.5), the depth z is defined as:

$$depth,\ z = \frac{fB}{d} \tag{3.7}$$

The equations used to calculate the exact location of $P_L(X,Y,Z)$ for the PTZ camera controller implementation in the next section are:

$$x = \frac{Bx_1}{d},\quad y = \frac{By_1}{d},\quad z = \frac{Bf}{d} \tag{3.8}$$

4. PTZ CAMERA CONTROLLER IMPLEMENTATION

The PTZ controller module is implemented in Matlab and using the ACTi Software Development Kit (SDK), which includes acquisition of the coordinates for its conversion to pan and tilt angles for the PTZ camera. A buffer of the coordinates is created for redundancy check and to improve the performance of the tracking system. The redundancy check block also helps preventing continuous unwanted execution, also saving computation time. The output from the face detection and stereo matching in the first subsystem is used to compute the pan, tilt and zoom angles.

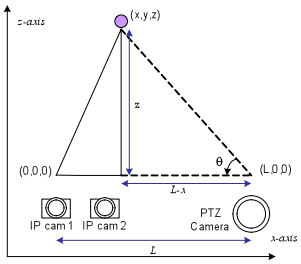


Figure 5. Pan angle computation

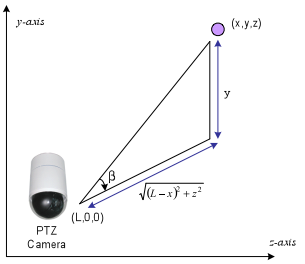


Figure 6. Tilt angle computation

The pan angle of the system is computed based on the model reported in Figure 5. The angle of the pan is defined as θ. The PTZ camera is placed at location (L,0,0) in the same X axis as the IP camera as illustrated in Figure 5. Thus, the angle is calculated as,

$$\theta = \tan^{-1}\left(\frac{z}{L-x}\right) \tag{4.1}$$

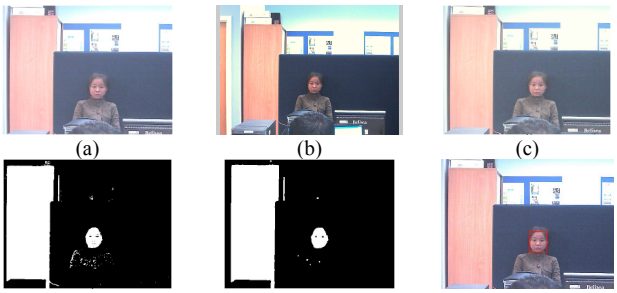
Meanwhile, the tilt angle β is computed based on the position of the PTZ camera at the same location (L,0,0), as shown in Figure 6.

$$\beta = \tan^{-1}\left(\frac{y}{\sqrt{(L-x)^2 + z^2}}\right) \quad (4.2)$$

The depth estimation is based on stereo matching algorithm with two IP cameras, enabling better image quality with the PTZ camera's zoom. The zoom ratio assigned depends on the distance of the object from the PTZ camera.

5. RESULTS AND DISCUSSION

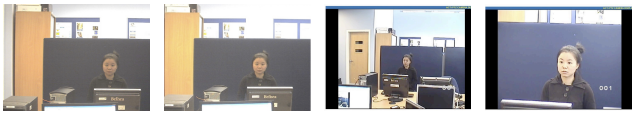
Figure 7 shows results of the face detection step. The original images are shown in Figure 7(a) and 7(b), for the left and right cameras respectively. The image, modified under adaptive illumination compensation algorithm, is shown in Figure 7(c). The algorithm removes overcast color lighting of the acquired images. Then it performs the skin color segmentation process. The skin color detection is implemented for different race skin colors. Some parts of the image may be identified as skin. For example objects in the background, as shown in Figure 7(d), where the wardrobe on the left side of the image is detected as "skin". Morphological processing is used to eliminate/reduce noise (Figure 7(e)). The face region is selected at the end of the process (Figure 7(f)), after computing the Euler number of each detected region.



(a) (b) (c) (d) (e) (f)

Figure 7. Results from video analytics module. (a) Left image; (b) Right image; (c) After adaptive illumination process; (d) Skin color segmentation; (e) Morphological processing; (f) Final result after connected region analysis.

The face detection result is processed in the block matching and location estimation step, 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's controller. The coordinates are converted into pan and tilt angles. Figure 8(a) and 8(b) show the images acquired by the left and right cameras respectively. With the stereo matching algorithm, the depth and location of the target object is evaluated and passed to the PTZ camera. The PTZ camera initially captures the targeted object as shown in Figure 8(c). Figure 8(d) illustrates a zoomed image of the object taken by the PTZ camera, with a zoom ratio of 5, at the end of all the described steps.



(a) (b) (c) (d)

Figure 8. Image taken by the intelligence surveillance system. (a) Image taken by left IP camera; (b) Image taken by right camera; (c) The PTZ captured the targeted object; (d) The object zoom by ratio of 5.

This 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 detection of the human face as object has been successfully implemented in the system with the +/-5% focal length correction error, neutralized by limiting the zoom of the PTZ camera to 30 times. The PTZ response upon changes of the coordinates has been found to be quick. Another major advantage of this system is that even with additional features, the whole system would be relatively cost effective for longer runs and could be used in real-time implementations. As Ethernet and LAN connections are almost prerequisite in all large industries, the system can be integrated easily with some calibration and IP address configuration in the initial setup.

6. CONCLUSION

A fully automated smart surveillance system has been designed and developed, being able to detect and zoom in on objects and acquire video data, using image processing techniques. The features of this system include face detection, high quality surveillance acquisition data using PTZ camera, and secure streaming of data due to password protection. The system consists of two fixed IP cameras and one PTZ camera. The system processes two images acquired from the fixed cameras, as a stereo input to calculate human face locations, which can be used to control the PTZ camera to find the object. Stereo matching algorithm is used to obtain correct corresponding block in the target image that is required for accurate location estimation. The final system allows the PTZ camera to track the object and acquire images in high-resolution quality.

7. REFERENCES

[1] M. Valera & S. A. Velastin, "Intelligent distributed surveillance systems: a review". *Vision, Image and Signal Processing, IEE Proceedings.* 192-204, 2005.

[2] L. D. Stefano, M. Marchionni & S. Mattoccia, "A Fast Area-based Stereo Matching Algorithm". *Proceedings from the 15th International Conference on Vision Interface* 22, 983-1005, 2004.

[3] A. Klaus, M. Sormann, & K. Karner, "Segment-Based Stereo Matching Using Belief Propagation and a Self-Adapting Dissimilarity Measure." *Proceedings of the 18th International Conference on Pattern Recognition* 3, 15 -18, 2006.

[4] S. Mattoccia, "A Locally Global Approach to Stereo Correspondence." *IEEE International Conference on Computer Vision Workshop, ICCV Workhops* , 1763-1770, 2009.

[5] D. Scharstein & R. Szeliski, "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondence Algorithms." *International Journal of Computer Vision* 47, 7-42, 2002.

[6] V. Vezhnevets, V. Sazonov & A. Andreeva, "A Survey on Pixel-Based Skin Color Detection Techniques." *Proc. Graphicon*, 2003.

[7] G. Buchsbaum, "A Spatial Processor Model for Object Colour Perception." *Journal of the Franklin Institute* 310, 1-26, 1980.

[8] J. Kovac, P. Peer, & F. Solina, "Illumination Independent Color-based Face Detection." *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis.*1, 510-515 Vol.1, 2003.

[9] A. Albiol, L. Torres & E. J. Delp, "Optimum Color Spaces for Skin Detection." *International Conference on Image Processing* 1, 122-124 Vol.1, 2001.

[10] A. Bovik, *Handbook of Image and Video Processing*. Elsevier Academic Press, 2005.

[11] Y. Morvan, "Acquisition, Compression and Rendering of Depth and Texture for Multi-view Video." *Thesis PhD*. Eindhoven University of Technology, 2009.