REAL-TIME OBJECT CORRESPONDENCE IN STEREO CAMERA SYSTEM

Fai Chan, Jiansheng Chen, and Yiu-Sang Moon

Department of Computer Science and Engineering,
The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{fchan, jschen, ysmoon}@cse.cuhk.edu.hk

ABSTRACT

In this paper, we address the problem of object correspondence construction in stereo camera systems by using a real-time algorithm adopting reverse stereo triangulation. This algorithm is based on a belief that any incorrect object-pair will eventually demonstrate inconsistency in its spatial location calculated from stereo triangulation, so that correct object-pairs can be identified from all possible object-pairs. We present experimental results from a dual camera human face capturing system in which more than 99% object correspondences can be accurately identified, while 100% of falsely detected objects are eliminated. Besides, our proposed method can handle no less than 100 object-pairs within 1ms in a P4 1.5GHz desktop PC.

Index Terms— Stereo vision

1. INTRODUCTION

The stereo matching task refers to the correspondence construction of each pixel in the image-pair captured by a stereo system. Such correspondences can serve as fundamentals for applications like 3D reconstruction and image stitching. However, in applications where only object correspondences are required, pixel correspondences [1, 2, 3] may not be necessary. Traditionally, the object correspondence problem is solved using geometry-based approach in which the epipolar geometry [4, 5] is used. With the knowledge of the fundamental matrix of the cameras, given a point in the image captured by one camera, the corresponding point in the other camera should be lying on an epipolar line. Nevertheless, if an active camera is used, its fundamental matrix may change according to its movement so that a costly re-computation of the fundamental matrix is required.

Existing pixel-based stereo matching approaches, working satisfactorily in 3D reconstruction, are generally too computationally expensive to be directly applied in object correspondence problems. In applications where object correspondence is sufficient, the computational resources have to be saved for other more urgent tasks. In this work, we address the object correspondence problem in a human face capturing system. One static webcam and one Pan-Tilt-Zoom (PTZ) camera are integrated into a stereo camera system. Once the human faces are detected and correctly corresponded, spatial location of a particular human face can be calculated. This information is then processed for controlling the PTZ camera to zoom and focus into the person for capturing a high resolution face image. This dual camera system may serve as the core of an active surveillance system capable of human face identification. To achieve this, human faces are first detected in both of the stereo images. Then, the problem is, given a set of possible face areas in each image of the image-pair, how can the correspondent human faces be identified while falsely detected face areas (false alarms) be filtered? This problem is challenging in two considerations: 1) this application requires real-time response and therefore accurate face correspondences is required to be performed within very short time durations; 2) face detectors are not perfect so that false alarms are inevitable and may jeopardize the reliability of the face correspondence process seriously.

In this paper, we propose a geometry-based algorithm for solving the face correspondence problem described above. The algorithm is based on the assumption that any non-correspondent object-pair will have inconsistencies in spatial locations calculated using stereo triangulation. Therefore, all the detected face pairs are processed using the stereo triangulation and only those with consistent spatial information will be regarded as genuine face-correspondence results. Experiments show more than 99% faces (detected in both images) are correctly corresponded and 100% of the false alarms from face detection are eliminated. Besides, the proposed method is fast enough to handle no less than 100 object-pairs within 1ms in a P4 1.5GHz desktop PC.

The organization of this paper is as follows. Section 2 introduces the stereo cameras system. In section 3, a skin-based human face detection algorithm is presented. The object correspondence algorithm is proposed in section 4. Experimental results are described in section 5.

2. A STEREO CAMERA SYSTEM

In this section, the stereo camera system used in this paper is described. The image pair captured by the system is used for locating the target human faces. Two distinct cameras are used in the system. A wide field webcam is used as the static (non-movable) camera for monitoring the whole scene and the other camera is an active PTZ camera.

2.1. Stereo Triangulation

The stereo camera system can be modeled using 2D trigonometry. Knowing the parameters, such as camera focal lengths or PTZ camera movement, spatial location of an object can be recovered by solving some trigonometric equations. This process is known as stereo triangulation.

A coordinate system is built as illustrated in Figure 2. Physically, the floor acts as a Ground Plane (X-Z). The plane containing the two cameras and is perpendicular to the Ground Plane is defined as Camera Plane (X-Y). Wall Plane (Y-Z) is perpendicular to both Camera and Ground Plane.

a) Camera Disparity T: distance between 2 cameras
b) Camera Height H: camera to ground height
c) Camera Pan Angles θ_1, θ_2: the angles between projection of camera direction vectors to the Wall Plane and the Z-axis.
d) Camera Tilt Angles ϕ_1, ϕ_2: the angles between projection of camera direction vectors to the Ground Plane and the Z-axis
e) Optical Points: $(x_c^1, y_c^1)(x_c^2, y_c^2)$
f) Focal Lengths: f_1, f_2

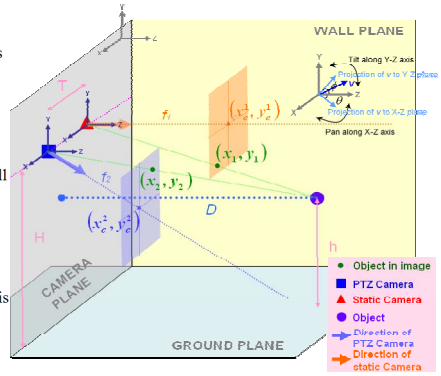

Figure 2: Illustration of stereo triangulation

The two cameras are set and assumed to have the same height from the ground; and both the static and the active camera are assumed to be directed to positive Z-axis initially. Suppose the coordinates of an object in the images taken by the two cameras are (x_1, y_1) and (x_2, y_2) respectively, as shown in the Figure 2. Formulae (1-3) can be derived using geometrical calculations. In the formulae, θ_1^i, θ_2^i and ϕ_1^i, ϕ_2^i are initial pan and tilt angle of the two cameras accordingly. D is the distance between the object and the Camera Plane. h_1 and h_2 are observed object heights from the two cameras.

$$\tan\alpha_1 = \frac{x_1 - x_c^1}{f_1} \quad \tan\alpha_2 = \frac{x_c^2 - x_2}{f_2} \quad \tan\beta_1 = \frac{y_1 - y_c^1}{f_1} \quad \tan\beta_2 = \frac{y_2 - y_c^2}{f_2} \quad \text{- (1)}$$

$$D = \frac{T}{\tan(\theta_1^i + \theta_1 + \alpha_1) + \tan(\theta_2^i + \theta_2 + \alpha_2)} \quad \text{--- (2)}$$

$$h_1 = H - D\tan(\phi_1^i + \phi_1 + \beta_1), \; h_2 = H - D\tan(\phi_2^i + \phi_2 + \beta_2) \quad \text{---(3)}$$

2.2. Camera system calibration

To estimate all the system parameters defined in the last section, a LSE (Least Square Estimate) based method is adopted. From Formulae (2-3), an optimization function can be constructed as Formula (4). The optimization function is simply the sum of square errors in calculating D for all the selected calibration points.

$$F(x_1, y_1, x_2, y_2, \theta_1, \theta_2, \phi_1, \phi_2, D, h) = \quad \text{--- (4)}$$
$$(D - \frac{T}{\tan(\theta_1^i + \theta_1 + \alpha_1) + \tan(\theta_2^i + \theta_2 + \alpha_2)})^2 +$$
$$(D - \frac{H - h}{\tan(\phi_1^i + \phi_1 + \beta_1)})^2 + (D - \frac{H - h}{\tan(\phi_2^i + \phi_2 + \beta_2)})^2$$

A set of correspondent points are manually selected as the calibration set. For each point, their height, object-to-camera-plane distance, camera pan and tilt angle are measured. Then, the optimization process is to find the parameter set such that the summation of the objective function (4) for all the selected points is minimal.

3. REAL-TIME HUMAN FACE DETECTION

As mentioned before, face detection will be performed for the images retrieved from both cameras in the stereo camera system. In our work, a skin-color model based on Y-C_b-C_r color space [6] is adopted for this task. It is based on the fact that human faces always look "pale-red" because of the capillaries on the surface of human skin. The constraint as illustrated in Formula (5) is imposed on C_r (Red-chroma) channel only. Figure 3 shows some examples.

(Y, C_b, C_r) is classified as skin if: $C_{r\_\min} \le C_r \le C_{r\_\max}$ --- (5)

(a) (b) (c)

Figure 3: a) Human Faces (with different skin colors), b) Corresponding skin maps and c) skin clustering results

After classifying each pixel into skin or non-skin, a skin map is produced as shown in Figure 3 b). Then human faces are extracted by clustering the large contiguous skin region by using the algorithm proposed by Synder and Cowart [7]. In the Figure 3 c), some corresponding clustering results of Figure 3 b) are illustrated.

4. OBJECT CORRESPONDENCE ALGORITHM

The proposed stereo camera system is deployed in a real life scenario where multiple faces may appear. As mentioned before, for a face candidate in the image of one camera, finding its correspondent face candidates in the image from the other camera is an object correspondence problem.

Furthermore, some of the detected face candidates may be false alarms, which should not have any correspondence in the counterpart image.

4.1. Information consistency of stereo triangulation

Recalling the stereo triangulation described in Section 2, given a pair of points in the stereo image pair, two observed object heights (from different cameras) h_1 and h_2 can be estimated. Although the observed heights are calculated from different cameras, their values should be very close to each other since they are actually estimates of the height of the same object. Based on this belief, a height consistency constraint can be formulated: *"If two points is correspondent, the observed heights from different cameras should be consistent and should not have a big difference."*

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

Figure 4: a) Left Image (One point is fixed), b) Right Image with all the points which satisfied Consistency Constraint with height difference set to 0.001m, c) 0.01m and d) 0.1m

This height consistency constraint can also be explicitly described by the epipolar geometry [5]. Figure 4 b) – d) are generated by finding all points satisfying the height consistency constraint with different tolerances for a fixed point as shown in Figure 4 a). In Figure 4 b), nearly exact height consistency is required and all the points are distributed along a line, or the epipolar line. Figure 4 c) – d) show the cases with more relaxed height difference tolerances. It is natural that the epipolar line expands to a region of which the area is related to the tolerances.

4.2. Proposed object correspondent algorithm

The proposed object correspondent algorithm is based on the observations described above. The notations and definitions in the algorithm are defined as follows.

a) c : Candidate Correspondent Pair Set
b) M_i : Candidate Correspondent Pair
c) M_{j_c} : Temporary Correspondent Pair Set
d) Dup_{M_i} : Number of conflicted correspondent pair of M_i
e) $Diff_{M_j}$: Difference of the observed height for the correspondent pair of M_i

The general idea of the proposed algorithm based on a belief that matched object-pairs always satisfy consistency constraint. The correspondence process is to select the correspondence pairs which are correctly matched.

Assume the objects found from image 1 are {A, B} and those from image 2 are {a, b}. Once a pair of correspondences is selected, say M_{Ab} (denoting the correspondence of Object A and Object b), any other correspondence pairs with Object A or Object b will never be chosen, since a correspondence of Object A and Object b has been established and those correspondences are said to be conflicting with M_{Ab}.

For a stereo image pair, firstly, a set of all possible correspondence pairs is built after object candidates are detected. By imposing the consistency constraint formulated in (6), most false alarms will be eliminated. Most remaining correspondences M_i from c are with target objects, but some of them may be in wrong orders because of their near locations. However, by observations, more than 95% of the incorrect correspondences are having more conflicts with those correct. Therefore, in the proposed algorithm, the correspondence bearing the largest number of conflicts is always eliminated. For correspondences having the same number of conflicts, the less "consistent" correspondence is removed. The operations iterate until there is no more conflict in the remaining set. The whole algorithm is illustrated in Figure 5.

Letting M_{12} to be corresponding pair with object P_1 and P_2 from the stereo image pair,
M_{12} satisfies the height consistency constraint if:

$$\left|h_1(P_1,P_2)-h_2(P_1,P_2)\right|<\delta \qquad \text{--- (6)}$$

0. **while** $Max_i(Dup_{M_i})>0$ do
1. **begin**
2. Compute Dup_{M_i} from c
3. Find correspondent set M_{j_c} s.t. $Dup_{M_j}=Max_i(Dup_{M_i})$ from c
4. **If** $count(M_{j_c})=1$
5. **then**
6. Remove all $M_j \in M_{j_c}$ from c
7. **else**
8. Select $M_j \in M_{j_c}$ s.t. $Diff_{M_j}=Max_i(Diff_{M_i})$
9. Remove selected M_j from c
10. **end**

Figure 5: Proposed Fast Object Corresponding Algorithm

5. EXPERIMENTS AND DISCUSSIONS

The purposes of the experiments are 1) to evaluate the proposed object correspondence algorithm, and 2) to investigate the effect of the algorithm to suppress false alarms. The proposed algorithm is implemented in the stereo camera system described in Section 2. The stereo camera system is deployed in an indoor office environment facing a 9-meter corridor. Experiments were performed on videos capturing two to three people walking towards the system. Each video frame is treated independently for face detection and face correspondences. The performance was evaluated using three video sequences. One of the video sequences contains two people and the others contain three people. In additions, all experiments were performed on a P4 1.5GHz desktop PC with 512MB RAM.

In the experiments, the resolution of the video captured by the webcam (Logitech QuickCam Pro 5000) and PTZ camera (AXIS 213 PTZ) are 640x480 and 704x576 respectively. The image dimensions are first down-sampled by four before applying skin detection and face clustering.

	Video 1	Video 2	Video 3	**Overall**
No. of video frames (x 2)	318	250	404	
1) ***Face detection rate*** = # correct face detected / # real face	0.8531	0.7858	0.8489	**0.8341**
2) ***Miss rate for face detection*** = 1 – *Face detection rate*	0.1469	0.2142	0.1511	**0.1659**
3) ***Successful rate for face detection*** = # correct face detected / # face detected	0.7502	0.921	0.8389	**0.831**
4) ***False alarm rate*** = 1 – *Successful rate for face detection*	0.2498	0.079	0.1611	**0.169**

Figure 6: Performance of Face Detection

According to Figure 6, the Face Detection Rate is 83.4% with False Alarm Rate of 16.9%. The performance of this face detector is acceptable considering the face areas in the video are always less than 50 x 50 pixels. The average computational time for face detection is 12ms for an image-pair. Figure 8 illustrates some of the detection results. The detected faces were surrounded by rectangles. The center points of these rectangles will be used for face object correspondences. There are some limitations to use this skin color model for face detection. Firstly, human limbs are sometimes misclassified as faces because of the presence of skin. Also, the model tends to misclassify objects with color in red or yellow as human face as shown in Figure 8. However, these false detection results can usually be filtered by the following object correspondence process.

	Video 1	Video 2	Video 3	**Overall**
No. of video frames	159	125	202	
1) ***Face correspondence Rate*** = # correct face correspondence / # real face detected in both images of stereo pair	1	0.984	0.9951	**0.9938**
2) ***Miss rate for face correspondence*** = 1 – *Face correspondence Rate*	0	0.016	0.0049	**0.0062**
3) ***Successful rate for face correspondence*** = # correct face correspondence / # face correspondence	1	1	1	**1**
4) ***False alarm rate*** = 1 – *Successful rate for face correspondence*	0	0	0	**0**

Figure 7: Performance of Face Correspondence construction

The detected faces in both images were then processed by the proposed object correspondence algorithm. The corresponding faces were constructed and surrounded by rectangles with same color in the image pair in Figure 8. According to the Figure 7, the Face Correspondence Rate is 99.4% and False Alarm Rate is 0%. Most false alarms were filtered by the height consistency constraint. Besides, Miss Rate for face correspondence is 0.62%. These results justified our observation that most incorrect corresponding rules are having more conflicts. The speed of the correspondence algorithm is further tested. Most image-pairs took less than 0.05 ms for the correspondence construction. For 100 object candidate pairs in simulation, it took less than 1ms for the whole correspondence process.

According to the experimental results, the proposed algorithm was found to perform satisfactorily for matching faces in stereo image pairs. Besides, it can suppress the false alarms produced by the face detection algorithm, so that relaxation in face detection algorithm becomes acceptable. Thus, simple and efficient face detection methods such as the skin color model used in this paper can be adopted despite of their high false alarm rate. More testing video sequences and the face correspondence results are available at: http://fpserver.cse.cuhk.edu.hk/icassp08_fchan.

6. CONCLUSION

In this paper, object correspondence problem in a dual camera human face capturing system is addressed. The face correspondences are constructed based on the belief that any incorrect object-pair will probably demonstrate inconsistency in object spatial locations calculated from the stereo triangulation. It is also observed that most incorrect correspondences have more conflicts than the correct correspondences. By imposing such constraints, 99.4% of the faces detected in both images are correctly corresponded with 100 matches per 1 ms on a P4 1.5GHz desktop PC. In the experiments, all the false alarms in face detection are successfully suppressed, so that simpler and faster face detection algorithms can be adopted to enhance the efficiency of the whole system.

This work was supported by the Hong Kong Research Grants Council Project 415207, "Biometric Authentication: from Security to Daily Life".

7. REFERENCES

[1] O. Veksler, "Fast Variable Window for Stereo Correspondence Using Integral Images", *Proc. of CVPR*, Vol. 1:556-561, 2003.

[2] N. M. Nasrabadi, Y. Liu, and J. L. Chiang, "Stereo vision correspondence using a multichannel graph matching technique", *Proc. IEEE Int. Conf. Robotics Automat*, 1988.

[3] P. Foggia, J.M. Jolion, A. Limongiello, and M. Vento, "A new approach for stereo matching in autonomous mobile robot application", *IJCAI*, pp. 2103–2108, 2007.

[4] Z. Zhang, "Determining the epipolar geometry and its uncertainty: A review", *IJCV*, 27(2):161–195, 1998.

[5] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, UK, 2000.

[6] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques", *GRAPHICON03*, pp. 85-92, 2003.

[7] W. Synder, and A. Cowart, "An iterative approach to region growing", *TPAMI*, 1983

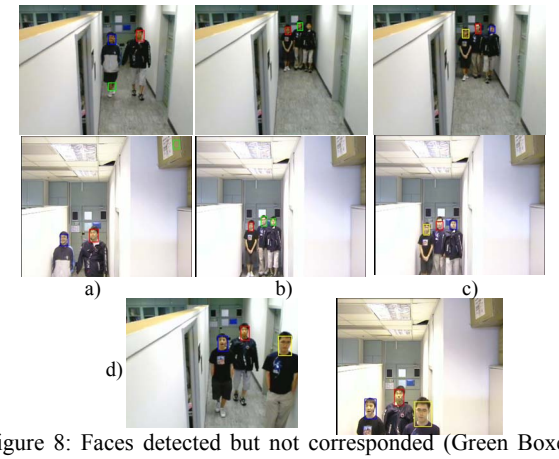

Figure 8: Faces detected but not corresponded (Green Boxes); but corresponded (Other colors). a) Video 1 Frame090, b) Video 2 Frame029, c) Video 2 Frame051, d) Video 3 Frame163