STEREO-BASED HUMAN HEAD DETECTION FROM CROWD SCENES

Xiaoyu Huang，*Liyuan Li*

Institute for Infocomm Research
A-Star, Singapore, 119613
Email: {xhuang,lyli}@i2r.a-star.edu.sg

Terence Sim

School of Computing, National University of Singapore
Singapore, 117543
Email: tsim@comp.nus.edu.sg

ABSTRACT

In this paper, a novel stereo-based head detection method is proposed for human detection in crowd scene. It contains three steps: (1) scale-adaptive filtering, (2) spurious clue suppression, and (3) human head location. With the depth information, the sizes of human heads could be estimated. From this, 3D scale-adaptive filtering is proposed. It is applied for extracting the likelihood evidence of heads from the stereo image. In the second step, the extracted points whose positions in the real space are much higher or lower than the average human height above the ground surface are further suppressed. Finally, human heads are located by applying a mean-shift algorithm to the likelihood map. Good results of detecting human heads in crowds have been obtained from the experiments on real scene.

1. INTRODUCTION

Detecting human individuals in a scene is one essential task for video surveillance. Commonly it is done by motion detection, e.g., background subtraction. However, when the scene is crowded with many overlapping persons in groups, it becomes difficult to detect human individuals from the foreground motion regions.

Velastin, *et al* [1, 2] proposed to use statistics of global image features, such as the texture, edges, and optical flows, to estimate the density of human objects in crowd scenes. Trained classifiers, e.g., neural network, are used to classify the scene into two to five categories, such as low, high, or very high of person densities. In the system [3], the density of persons is estimated by counting the foreground pixels with the weights based on the perspective correction. Zhao and Nevatia proposed a method to segment human individuals in foreground motion regions [4]. They try to interpret the foreground region with a configuration of a number of human individuals by maximizing the posterior probability. Specific knowledge about human shape, height, camera setting and image cues of head contours are integrated in a Bayesian framework. Thousands of iterations are required to get an optimal solution for a frame containing many persons. Meanwhile, existing stereo-based method to track isolated human object close to the camera is not applicable to detect human heads in crowds [5]. In general, it is still very difficult to detect human individuals in crowds from 2D motion regions involving many overlapping people.

For a surveillance system to monitor a busy public site, e.g., the airport, railway station, bank, or shopping center, the camera usually looks the scene from a high position. Since every human being occupies a 3D volume on the ground surface, human heads are isolated from each other in 3D space even in crowds. Motivated from this observation, we propose a method to detect human heads in crowds from stereo images. The main contributions of this paper are: (1) propose a scale-adaptive filtering approach to extract the evidence for head like objects from the stereo image; (2) propose a method of restoring the perspective of the view to suppress spurious clues which are much higher or lower than the average human height above the ground surface; (3) propose a mean-shift algorithm to detect and locate human heads in the likelihood map. Promising results have shown that this method is applicable to crowd scenes.

The remaining of this paper is organized as follows. In Section 2, the scale-adaptive filtering is described. Section 3 discusses suppressing spurious clues based on the perspective of the view. In Section 4, a mean-shift algorithm to locate human heads in the likelihood map is presented. Finally, the experimental results and conclusions are given in Section 5 and 6, respectively.

2. SCALE-ADAPTIVE FILTERING

The stereo image is obtained with a wide-baseline stereo camera from SRI's Small Vision System [6]. It can generate the disparities for objects within 50m. The relation between the disparity (d) and the depth (z) to the camera is $d = bf / z = K_1 / z$, where b is the baseline, and

0-7803-8554-3/04/$20.00 ©2004 IEEE.

f is the focal length of the lens. $K_1 = bf$ is a constant for each stereo camera. After calibration, the disparity image $g(x,y)$ can be obtained from the left and right images captured by the stereo camera. One example of the color and disparity images are shown in Fig. 2 (a) and (b).

Since the stereo information allows us to estimate the size of a human head with the corresponding distance to the camera, scale-adaptive filters could be designed to aggregate the stereo evidence for a head and suppress that of other objects. Let $d = g(x,y)$ be the disparity value at the pixel (x,y). If it is the center of a head with the distance $z = K_1 / d$ to the camera, the disparity values from the head will be within the range of $[d_-, d_+] (d_- < d < d_+)$ with $d_\pm = K_1/(z \mp D_h/2) = K_1 d/(K_1 \mp D_h/2)$, where D_h is the average depth of human heads. And more, from the simple perspective triangles, the estimated width and height of the head in the image can be obtained as

$$w_h(d) = fW_h / z = K_2 W_h / d, \quad h_h(d) = K_2 H_h / d, \tag{1}$$

where $K_2 = fK_1$ is a constant, w_h and h_h are the average width and height of human heads, respectively. Now we know that if *(x,y)* is the center of a human head, the cloud of the disparity data from this head will be distributed within a 3D bounding box centered at it with width, height, and depth range being w_h, h_h and $[d_-, d_+]$, respectively. Besides, since the human heads look like the isolated balls from an elevated vantage point, there is no object on the top and both sides of it with the same depth distance to the camera. This observation could be used to suppress the disparity measures for other objects, such as human bodies.

From the estimated 3D sizes of the head, the filters for scale-adaptive filtering can be designed as follows. First, a 2D window is generated for a possible head centered at *(x,y)* as shown in Fig. 1 (a). The size of the window is

$$W_w^{(\gamma)} = 2\gamma w_h(d), \; H_w^{(\gamma)} = 1.5\gamma h_h(d) \tag{2}$$

where γ is a scale factor to adapt to the scale variations for different persons. Here $\gamma \in \{0.8,1,1.2\}$ are used. With the layout as shown in Fig. 1 (a), the possible pixels belonging to the head would be within a ellipse approximating the head contour. Let the distance of a point (x',y') within the window to the head center be

$$d_\gamma(x',y') = (x'-x)^2/a_\gamma^2 + (y'-y)^2/b_\gamma^2 \tag{3}$$

where $a_\gamma = \gamma w_h(d)/2$ and $b_\gamma = \gamma h_h(d)/2$. Then a 2D spatial filter is defined as

$$F_S^{(\gamma)}(x',y') = \begin{cases} 1, & d_\gamma \in [0, 0.7] \\ 2(1 - d_\gamma / 0.6), & d_\gamma \in (0.7, 1.3] \\ -1, & d_\gamma > 1.3 \end{cases} \tag{4}$$

where d_γ denotes $d_\gamma(x',y')$. In practice, $F_S^{(\gamma)}(x',y')$ is a weight mask which would give positive supports for the evidence within the ellipse and negative supports for that out of the ellipse. Because we are only interested in the points in the window whose disparity values are within $[d_-, d_+]$ for the possible head centered a (x,y), the depth filter is defined as

$$F_D(x',y') = \begin{cases} [g(x',y') - d_-]/(d - d_-), & g(x',y') \in [d_-, d) \\ [d_+ - g(x',y')]/(d_+ - d,) & g(x',y') \in [d, d_+] \\ 0, & otherwise \end{cases} \tag{5}$$

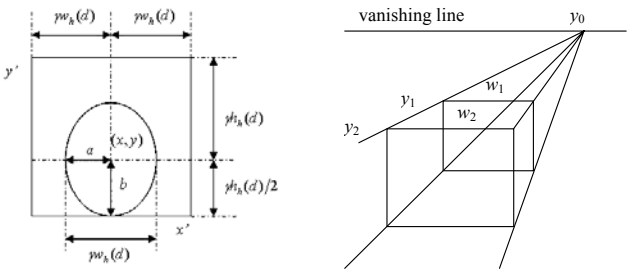
$F_D(x',y')$ gives a larger weight for a disparity value closer to d. With the spatial and depth filters for the pixel *(x,y)*, the result of scale-adaptive filtering with the scale factor γ is defined as

$$e_\gamma(x,y) = \frac{\alpha}{W_w^{(\gamma)} H_w^{(\gamma)}} \sum_{(x',y')} F_S^{(\gamma)}(x',y') F_D(x',y') \tag{6}$$

where *(x',y')* are the pixels within the window, and α is a constant for normalization. For α=3.3, the output is within [0,1]. From (4), (5) and (6), it can be seen that, for the pixels *(x',y')* whose disparity values are within $[d_-, d_+]$, if they are within the ellipse they will give positive support for *(x,y)* belonging to a head, otherwise, they will give negative evidence of *(x,y)* being a part of a head. To adapt to the scale variations, the likelihood map generated by scale-adaptive filtering is defined as

$$e(x,y) = \max_\gamma \{e_\gamma(x,y)\} \qquad \gamma \in \{0.8,1,1.2\} \tag{7}$$

In *e(x,y)*, there are several bright blobs which correspond to possible heads. The example of *e(x,y)* for Fig. 2 (b) is shown in Fig. 2 (c).

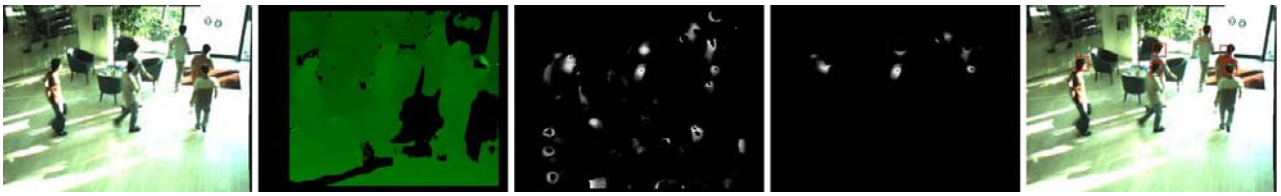


(a) (b)

Fig. 1. (a) Spatial layout of the scale-adaptive filter. (b) The perspective of a view from an elevated vantage point.

3. SUPPRESSION OF SPURIOUS CLUES

In the likelihood map *e(x,y)*, there would be some bright blobs not for human heads. They might be generated by human body parts, background objects, and shadows on the ground surface. To filter out such spurious evidence for human heads, a virtual plane which is parallel to and above the ground surface with average human height is established. The points in *e(x,y)* are compared with the plane in real space and those which are far away to the plane are suppressed. The details are



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

Fig. 2. Example images for head detection: (a) the color image; (b) the disparity image; (c) the likelihood map after scale-adaptive filtering; (d) the likelihood map after suppressing spurious clues; (e) the detected heads.

described as follows.

In an up-right view from an elevated vantage point, if there are two rectangles of the same size standing on the ground surface with different depth distances to the camera, the perspective geometry can be illustrated by Fig. 1 (b). The lines connecting corresponding corners of the two rectangles will meet at a horizontal line, the vanishing line. Let y_1 and y_2 be the vertical positions of the tops, w_1 and w_2 be the widths of the two rectangles in the image, and y_0 be the vertical position of the vanishing line. Then, from the perspective geometry and (1), we have $(y_2-y_0)/(y_1-y_0)=w_2/w_1=d_2/d_1$, where d_1 and d_2 are the corresponding disparity values from the two top-left corners. Let's suppose the height of the rectangles is the average human height H_p, y^* is the position of a rectangle's top-left corner, and d^* is the disparity value of it. Then, if there is another rectangle with y and d for its top left corner from the disparity image, y can be solved, which gives a virtual plane as $y=Ad+B$, where $A=(y^*-y_0)/d^*$ and $B=y_0$ are two constants. The virtual plane is parallel to and above the ground surface with the height of H_p. This virtual plane can be established in an initialization step. Let's capture several sample images with different persons in different positions in the scene, and manually mark the head positions in the images. Then we can get a set of training samples. By using Least-Square fitting, we can obtain the estimated plane $y = \hat{A}d + \hat{B}$.

In the process of head detection, each pixel of *e(x,y)* is scanned. If *e(x,y)* > 0, from (1), the distance of the point in the real space to the virtual plane is calculated as

$$\Delta H(x,y) = \frac{|\, y - (\hat{A}g(x,y) + \hat{B})\,|}{K_2 g(x,y)} \tag{8}$$

A weight of *g(x,y)* belonging to a head is then generated as

$$r_h(x,y) = \begin{cases} 1, & if\ \Delta H \le 0.15m \\ (0.3 - \Delta H)/0.15, & elseif\ \Delta H \le 0.3m \\ 0, & otherwise \end{cases} \tag{9}$$

where ΔH is $\Delta H(x,y)$. Now the suppression of the spurious evidence is performed as

$$\tilde{e}(x,y) = r_h(x,y)e(x,y) \tag{10}$$

The example of the final likelihood map $\tilde{e}(x,y)$ for Fig. 2 (b) is shown in Fig. 2 (d). It can be seen that most of the spurious blobs have been suppressed and just the significant blobs for real human heads are remained.

4. HEAD DETECTION

The last step for head detection is locating the bright blobs in $\tilde{e}(x,y)$. The heads are extracted one by one. For each human head, the following operations are performed.

First, scan the map $\tilde{e}(x,y)$. If the maximum value $\tilde{e}(x,y)$ >0.2, it is set as the initial seed to locate the head. Let *d*=*g*(*x*,*y*), we can obtain $w_h(d)$ *and* $h_h(d)$ by (1). An initial window B_0 with size of $w_h(d)\times h_h(d)$ and centered at $(x_0,y_0)=(x,y)$ is established.

Secondly, a mean-shift algorithm [7] is applied to locate the head iteratively. At each iteration, the new center of the head is calculated as

$$s_t = \sum_{(x,y)\in B_{t-1}} s\tilde{e}(x,y) / \sum_{(x,y)\in B_{t-1}} \tilde{e}(x,y),\ s = x \text{ or } y \tag{11}$$

Now the window center is moved from (x_{t-1},y_{t-1}) to (x_t,y_t). If $|x_t-x_{t-1}|<3$ and $|y_t-y_{t-1}|<3$ or $t>10$, the mean-shift algorithm is terminated. The position (x_t,y_t) is the detected center and B_t is the bounding box of the head. If the evidence from the window is too small, the blob is eliminated.

Finally, set $\tilde{e}(x,y)=0$ for the pixels within the window B_t. Then start to find next human heads in $\tilde{e}(x,y)$.

If no $\tilde{e}(x,y)$ > 0.2 is found in a scan, the process of head detection is finished. For the detected blobs, if one is under another of larger evidence and the horizontal distance between them is less than 0.2m, it might be the shoulder and suppressed further. The remainders are the detected heads. The example of the head detection from

Fig. 2 (b) is shown in Fig. 2 (e) where the detected heads are bounded with red windows and the centers are marked with red points.

5. EXPERIMENTAL RESULTS

The proposed method has been tested on the real images captured in an entrance hall of an office building. The test images contain various scenes of crowds, including groups of people walking together or standing and talking to each other. We also captured the images from different view points. One example of 6 persons walking in the hall has been shown in Fig 2. Another example of 6 persons standing and talking to each other is shown in Fig. 3 (a). In these two examples, there are many cases of two persons overlapping in the depth direction, which are the most difficult cases for the methods on the 2D regions [4]. The proposed method detects the persons correctly for these tough cases. One more example of 10 persons walking together is shown in Fig. 3 (b). In the example images, there are some misalignments of head locations. This is because the disparity images are generated after warp rectification and by a region-based method [6].

(a) (b)

Fig. 3. Two more examples of head detection.

Table 1. Systematic Evaluation for Head Detection.

	Number	Rate
correct detections	$N_d = 503$	$N_d / N_p = 91.62\%$
false detections	$N_f = 48$	$N_f /(N_d + N_f) = 8.71\%$

To systematically evaluate the performance of the method, the results from 195 images (including those with poor illuminations) are compared with ground truth by hand. The evaluation results are summarized in Table 1, where $N_p = 549$ is the number of valid persons (ground truth). The detection rate indicates how many truth persons are detected with respect to the ground truth, and the *error rate* shows percentage of false detections in the total detected persons. The statistics show that promising results have been achieved by the proposed method.

Another advantage of the proposed method is that it does not depend on background subtraction. So it is less sensitive to the shadows and other background changes.

6. CONCLUSIONS

In this paper, a stereo-based solution to the problem of detecting human individuals in crowds is proposed. The depth information allows us to estimate the size of human heads in the image. Based on such estimation, a process of scale-adaptive filtering is proposed to extract the evidence for the presence of human heads while suppress that of other objects. To suppress the spurious clues further, a virtual plane which is parallel to and above the ground surface with average human height is established. The extracted clues whose 3D positions in the real space are far away to the virtual plane are suppressed. Finally, a mean-shift algorithm is proposed to locate the bright blobs in the likelihood map for human head detection. Promising results have obtained from the experiments on real scene.

Besides stereo (disparity) information, motion cues could also provide the information of human individuals in the scene. We intend to involve the motion cues for both detection and tracking of human heads in future.

7. REFERENCES

[1] A.C. Davies, J.H. Yin, and S.A. Velastin, "Crowd monitoring using image processing," *Electronics & Communication Engineering Journal*, pp. 37–47, Feb. 1995.

[2] A.N. Marana, S.A. Velastin, L.F. Costa, and R.A. Lotufo, "Automatic estimation of crowd density using texture," *Safety Science*, vol. 28, no. 3, pp. 165–175, 1998.

[3] N. Paragios and V. Ramesh, "A mrf-based approach for real-time subway monitoring," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. I–1034–I–1040, 2001.

[4] T. Zhao and R. Nevatia, "Bayesian human segmentation in crowded situations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* 2003.

[5] D. Russakoff and M. Herman, "Head tracking using stereo," *Proc. IEEE Workshop on Applications of Computer Vision,* pp. 254–260, 2000.

[6] K. Konolige, "Small vision system: Hardware and implementation," *Proc. Int'l Symp. Robotics Research,* pp. 111–116, 1997.

[7] D. Comaniciu and P. Meer, "Mean shift analysis and applications," *Proc. Int'l Conf. Computer Vision,* pp. 1197–1203, 1999.