Stereo Head/Face Detection and Tracking

Jian-Gang Wang, Eng Thiam Lim and Ronda Venkateswarlu
Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
jgwang@i2r.a-star.edu.sg

Abstract

This paper presents a real-time head/face tracking system that uses synchronized disparity and intensity frames from stereo cameras. It integrates stereo, morphological watersheds, gradient and feature extraction techniques to track the head/face. The face is tracked once the facial features, e.g. eyebrows, eyes, mouth and nostrils, are available. The head is tracked if the facial features are not available, e.g. when the person is away from the stereo head or when the head is in the profile view. Our approach automatically initializes without user intervention and can be re-initialized automatically if tracking of the 3D face pose is lost. The results on our face database and real image video sequences illustrate the effectiveness of the method.

1. Introduction

There are few approaches on stereo head detection and tracking. Darrell et al [4] integrated stereo, color and pattern recognition to track the head but heavily relied on the skin-tone pixel extraction. Huang et al [6] used only stereo but the performance is satisfactory only when the human motion is slow and manual initialization is required. Luo et al [8] used stereo to detect heads by subtracting the background, for which prior knowledge of background is required. Very recently, Daniel et al [5] presented a head/torso tracking method using only stereo. However, the background has to be modeled and updated. Yang et al [9] presented a model-based stereo head tracking system. An interactive selection of seven landmark points in each face image is required to estimate the initial head pose.

In this paper, we propose a hybrid approach to detect and track head/face in real-time. Combining disparity and intensity image, either head or face/3D pose is tracked automatically. The head is tracked if face features are not available, e.g. when the person is far away from stereo head or with the profile view; face and 3D pose are tracked once the facial features, e.g. and nostrils, eyebrows, eyes, mouth are available. Disparity map of the face is obtained at frame rate by commercially available stereo software, e.g. SRI International [7] Small Vision System (SVS). The range data of a person is extracted in the disparity map by assuming the person of interest is the nearest object to the camera. A novel method is proposed where the head is separated from the connected head-and-shoulder component using morphological watersheds. The eyebrows, eyes, mouth and nostrils are darker than the facial skin, hence the features within the ellipse are extracted using median filter in our approach. The head contour is modeled as an ellipse, which can be least-squares, fitted in the watershed segmentation image. The face contour is also modeled as an ellipse. This ellipse can be fitted to the face by maximizing the mean gradient along the perimeter of the ellipse. Using the calibrated parameters of the vision system, the head pose can be estimated using the face plane formed by the features. The signal flow diagram is shown in Fig. 1.

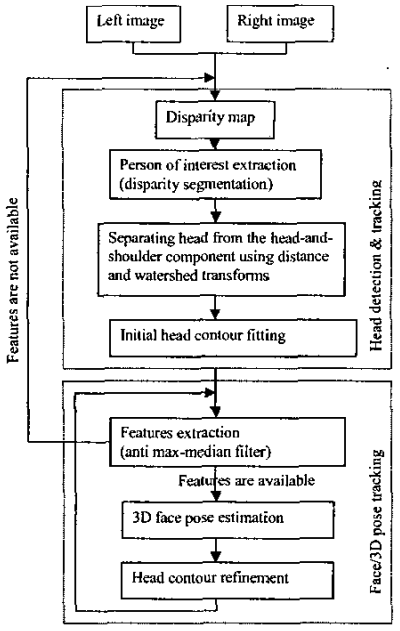


Figure 1 Flowchart of the head/face-tracking algorithm

2. Head detection and feature extraction

2.1 Object-oriented disparity segmentation

In this paper, we would like to track only the nearest face to the camera. Since we are using stereo system, we can separate the nearest face from other faces by alysis of the disparity histogram. This will help tracking the objects efficiently. This is not possible in case of a single camera system. Two people in front of the camera are separated using the disparity map as shown in Figure 2.

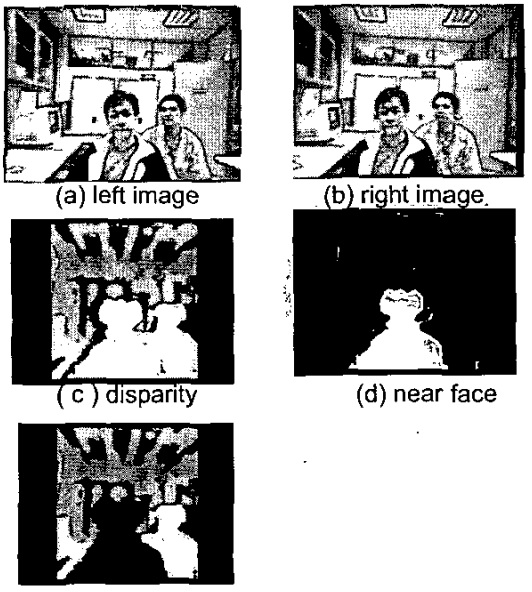


Figure 2 Separating faces using disparity histogram

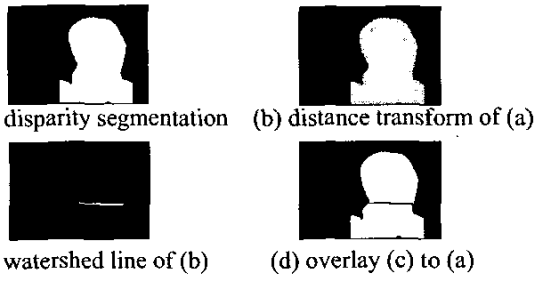


Figure 3 Separating the head from the shoulder using morphological watersheds

2.2 Head location using morphological watersheds

A novel method, which employs morphological watersheds transform in conjunction with distance transform to separate head from shoulder, is proposed in this paper. Watershed is an efficient tool to detect touching objects. To minimize the number of valleys thresholding their distances from the stereo head. The thresholds are selected based on the peak an

found by the watershed transform, one need to maximize the contrast of our objects of interest. Distance transform determines the shortest distance between each blob pixel and the blob's background, and assigns this distance to the pixel. Here, we use distance transform to produce a maximum in the head and shoulder blobs, respectively. In addition, the head and the shoulder have touching zones of influence. Applying watershed to the resulting distance transformation, the head is located using the watershed line. The segmented head contour is least-squares fitted as an ellipse. We have tested our algorithm on a face database built in our lab. There are 1000 face images of 100 student volunteers with 10 different head poses. At each pose, the database includes two intensity face images (stereo) and one disparity image computed using SVS. The experimental results are satisfactory. An example is shown in Fig. 3. Some of the resulting frames of the sequences of a person are shown in Figure 4. Frame 2, 20, 40, 60 and 80 are shown respectively from top to down.

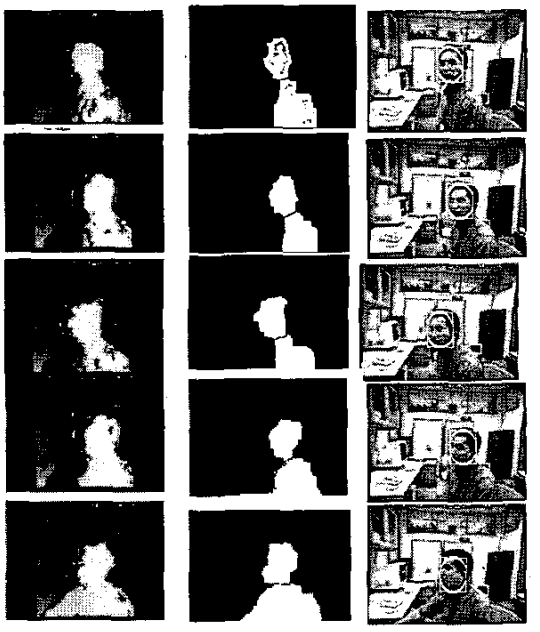


Figure 4 Watershed segmentation and face contour location on sequence of real face images; left column: disparity; middle column: watershed segmentation; right column: elliptical contour

Although the algorithm works well, the watershed line may not incidentally correspond to the neck. Fortunately, this can be detected by checking the ratio of the output elliptically head. The ratio of the head contour is assumed

1.2 in our experiment. If the watershed is found not correspond to the neck, i.e. the difference between the ratio of the output head and the assumed one is significant, a candidate head will be placed below the vertically maxima of the silhouette, in a manner similar to Darrel [4], and will be refined in the feature tracking stage.

2.3 Feature extraction

In this paper, we propose a new face feature extraction method, called anti-max median filter, which the features including eyes eyebrows, nostrils and mouth are detected by subtracting the max-median filter image from the original image. The knowledge, namely that the eyes, eyebrows, nostrils and mouth are darker in the gray level image, is the reason that the anti max-median filter would work. Although median filter is normally a somewhat slower process than convolution, due to the requirement for sorting all the pixels in each neighborhood by gray level, we adopt it because only the pixels in the face region are filtered. There are, also, algorithms that speed up the process [1].

The gray level image within the ellipse is subjected to the anti-max-median filter to extract face features as shown in Fig 6. We look for the features using the major axis of the ellipse, e.g. two eyes should be at the two sides of the axis, nostrils should keep nearly same distance from the two eyes. These features are mainly used for tracking purpose. In our implementation, image is sampled to 160 × 120 in order to speed up the median filter, size of 11×11 filter is used. On the other hand, the anti-max-median filter was applied to the face region instead of the whole image in order to save time.

(a)Original;(b) Max-median filter of (a);(c)=(a)-(b)

Figure 6 Feature extraction

3. Face pose tracking

The 2D positions of the facial feature and the ellipse contour are used to track the face simultaneously. In order to track face robustly even the illumination is changed, the areas around the detected features are tracked by mean subtracted normalized correlation, which is accurate and can take care of illumination changes. This computation is affordable because the search region is restricted to within the elliptical face contour.

The positions of the features in a frame are predicated from the previous two frames:

$$\mathbf{x}(t)=\mathbf{x}(t-1)+c(\mathbf{x}(t-1)-\mathbf{x}(t-2)) \quad (1)$$

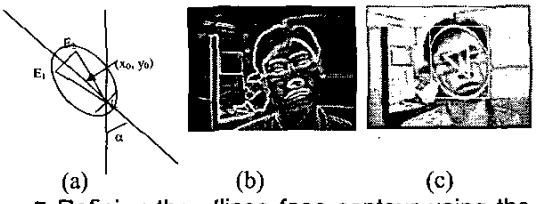
where $\mathbf{x}_{-1}=\mathbf{x}_0$, c is a constant.

The face is modeled as a vertical ellipse. Different from [2], we consider α, angle between the symmetry axis of the face (connect the center of the two nostrils and the center of the two eyebrows) and the vertical direction, see Fig. 7(a). Considering α, the face contour corresponds to an ellipse (x_0, y_0, b, α) where the normalized sum of the gradient magnitude along the perimeter of the ellipse is maximum:

$$G(x,y,\alpha)=\max_{b}\left(\frac{1}{N}\sum_{i=1}^{N}\left|\mathbf{n}(i)\cdot g(x_0,y_0,b,\alpha)\right|\right) \quad (2)$$

where $\mathbf{n}(i)$ is the unit vector normal to the ellipse at pixel i, g is the intensity gradient at pixel i, and · denotes dot product. The ellipse (x_0, y_0, b, α) centers at (x_0, y_0), direction of the major axis is with vertical angle α and the length of the minor axis is b. The ratio of the ellipse is assumed 1.2 units.

Using the above constraints, including the direction of the ellipse major axis and the ellipse center, the contour of the face instead of the contour of the head is tracked when the face features are available. Initializing and re-initializing, which are done automatically.



(a) (b) (c)

Figure 7 Refining the ellipse face contour using the facial features (a) facial features, center and the direction of the major axis of the ellipse; (b) gradient image of Fig. 6(a); (c) feature triangle and the refined ellipse.

The features could be occluded due to the turning of the head. Once this happens, the head detection module will start automatically. That means the contour of the head instead of the contour of the face is tracked if the facial features are not available. If there is not enough of the size of the region in the disparity segmentation image, the system will report there is no person with the depth range of interest

4. Experimental results

We have evaluated our algorithm on video sequences of stereo image as the person is moving in various directions in front of a PC. The algorithm runs on a Pentium III 733M Hz PC in a speed about 25 frames per

second. SVS stereo heads MEGA-D MegaPixel' is used for implementing the system. SVS runs at rates of up to 45Hz with image resolution 320×240, and gives for each pixel the disparity between the images from stereo head. The calibration and the rectification of the cameras are done automatically using SVS library.

Feature tracking are shown in Fig. 8. The features, including two centers of the two eyes, the center of the nostrils, are tracked. Pose is estimated using these features and results are shown in Fig. 9.

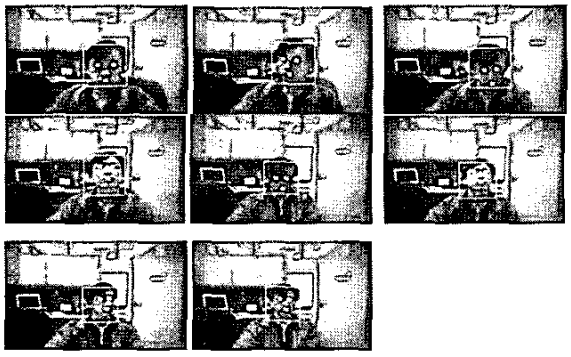


Figure 8 3D pose tracking

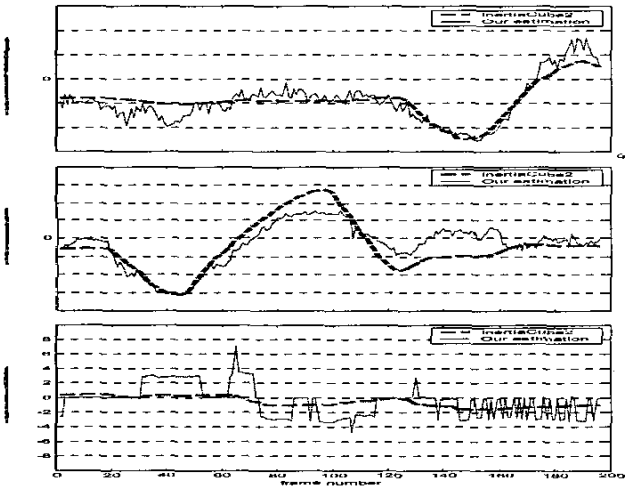


Figure 9 Comparison of our pose estimation result with InertiaCube2. From top to bottom are the rotation angles around X, Y, and Z axis. The RMS errors are 2.92, 5.70 and 1.46 degree respectively.

5. Conclusions and future work

The contributions of the paper include head segmentation using morphological watersheds, features extraction using anti max-median filter and refinement of face contour with facial features using improved gradient ellipse model. This method facilitates automatic initialization. One of our motivations here is to develop a tracking (video) based face recognition system. Video-based face recognition offers several advantages over still-image-based face recognition in the applications such as access control and ATM, where the video is acquired in a relatively controlled environment and the face is also relatively large. Video provides abundant image data, hence we can select good frames on which to perform classification; video allows tracking of face images, hence phenomena such as facial expressions and pose changes can be compensated for, resulting in improved recognition.

6. References

[1] J. T. Astola and T.G. Campbell, "On computation of the running median," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 37, No. 4, pp.572-574, 1989.

[2] S. Birchfield, "An elliptical head tracking using intensity gradients and color histograms," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, Santa Barbara, California, pp.232-237, June 1998.

[3] T.Choundhury, B. Clarkson, T.Jebara and A.Pentland, "Multimodal Person Recognition using unconstrained audio and video," in *Proceedings of International Conference on Audio- and Video-based Person Authentication*, pp.176-181, 1999.

[4] T. Darrell, G. Gordon, J. Woodfill, M. Harville, "Integrated person tracking using stereo, color and pattern detection," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*, pp.601-609, San Barara, Calif, June 1998.

[5] B. Deniel and M. Herman, "Head tracking using stereo," *Machine Vision and Applications*, Vol. 13, pp.164-173, 2002.

[6] N. Jojic, M.Turk and T. Huang, "Tracking self-occlude articulated objects in dense disparity maps," in *Proceedings of IEEE International Conference on Computer Vision*, pp.123-130, 1999.

[7] K. Konolige, "Small Vision Systems: Hardware and Implementation," in *Proceedings of Eighth International Symposium on Robotics Research*, Hayama, Japan, October, 1997. http://www.ai.sri.com/~konolige/

[8] R. Luo and Y. Guo, "Real-time stereo tracking of multiple moving heads," in *Proceedings of International Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, pp. 62-67, 2001.

[9] R. Yang and Z. Zhang, "Model-based head tracking with stereo vision," in *Proceedings of the fifth International Conference on Automatic Face and Gesture Recognition*, pp. 255-260, 2002.