

Head Segmentation and Head Orientation in 3D space for Pose Estimation of Multiple People *

Sangho Park and J. K. Aggarwal

Computer and Vision Research Center

Department of Electrical and Computer Engineering

The University of Texas at Austin, Austin, Texas 78712, USA

{ sh.park, aggarwaljk } @mail.utexas.edu

Abstract

We present an algorithm for establishing head orientations of multiple persons in 3D space. Using multiple features from grayscale images (i.e. binary blobs, silhouette contours, and intensity distributions), our algorithm achieves foreground separation, head segmentation, and head-orientation classification, respectively. The information is then combined to form an integrated representation about how the heads of multiple persons are configured in 3D space in order to describe their relative position. The algorithm classifies each head orientation, ranging from 0 to 360 degrees of rotation on a horizontal plane, into eight classes by using a moments-based method. The algorithm can be easily extended to video sequences of image frames for describing how head poses change over time in relation to each person involved in a scene. Experimental results are presented and illustrated.

1. Introduction

Interpretation of human behaviors in video data is essential in human-involved applications such as surveillance, human-motion interpretation, and human-computer interface. Visual interpretation of humans is a challenging task because it involves detection of the human form, which is subject to deformation, recognition of articulated motions, and understanding of interactions between people. Due to the complexity of these tasks [1], most studies have concentrated on some subset of particular body features to input for visual interpretation. The face and head are among the most commonly used body parts.

Information from the human head image is very useful, as it can enable us to recognize faces, interpret emotions, identify a particular person, and track head motion. Most studies on the human image have been restricted to a single person in isolation. Study of interrelationships between multiple persons is very new.

Our overall motive is to build a recognition system that interprets the spatial configuration of multiple interacting persons. In a multiple-person situation, it is important to interpret relative relations between people in terms of space and time so that a vision system can ‘understand’ how people behave toward each other and what is going on in a given scene. Understanding of relative behaviors between people makes it necessary to understand images from an object-centered viewpoint as well as from a camera-centered viewpoint. For example, a person may stand facing another person instead of facing the camera and their gestures would be directed at that person instead of at the camera. Therefore, we need to consider all viewing directions from the camera, including the front-view, profile-view, and rear-view of human body.

Existing studies related to head detection or head orientation classification are limited in various respects. Haritaoglu et al. [5] used histogram projection of a silhouette image to detect multiple persons’ heads, representing each head as a single point in 2D space. Horprasert et al. [6] detected 3D head orientation from a monocular image sequence by applying a 3D to 2D projection of facial feature locations such as eyes and nose tip. Their algorithm is restricted to front-view and half profile view of faces because their algorithm depends upon locating these facial features. Baluja [2] presented a system that can cover the orientations ranging from left profile view to right profile view. But his algorithm requires manual normalization of image size. Darrell et al. [3] presented a system that tracks a face and estimates the facial pose. Their algorithm is view-based and does not depend on facial features, but it is also restricted in the viewing range of the face images

* This work was supported in part by the Army Research Office under contracts DAAG55-98-1-0230 and DAAD19-99-1-0012 (Johns Hopkins University subcontract agreement 8905-48168) and the State of Texas Higher Education Coordinating Board Advanced Research Project 97-ARP-275.

and requires a number of different sets of eigen-spaces to cope with different image sizes. Shirohey[9] applied an ellipse model to edge maps of an image in order to achieve face segmentation and identification. But his algorithm is applied only to front-view head images. Most of the studies are restricted in viewing directions (i.e., profile-view or front-view of face), require a normalized image size, or depend on facial features' relative locations in a face image which requires high-resolution and reliable detection of facial features.

In this paper we present a new approach to head segmentation, head-orientation classification, and configuration of multiple heads in 3D space. Our goals include detection of multiple head orientations ranging from 0 to 360 degrees of horizontal rotation, image size-independent processing, and configuration of multiple heads in 3D space using stereo. Without 3D information, the head orientation derived only with a 2D head image would cause ambiguity in establishing the relative relationship of visual attention indexed by head orientation, because there is no information about the location of the heads in 3D space. Therefore, we need to configure the spatial positions of the heads in 3D space in order to know whether people in a scene can actually see each other. The stereo algorithm also provides information about the actual proximity between people. Our goals make it necessary to combine multiple processing results using binary blobs, silhouette contours, and intensity information in the image in order to interpret the relative relations between people.

Figure 1 shows a diagram of the proposed system. Inspired by biological vision [7], this system is based on multiple processings which are independent of each other. We first acquire the background image and the input image. In preprocessing, we extract the binary blobs, the contour of the silhouette, and the intensity of the foreground areas. Each of these is processed by the stereo-vision module (S.V.), head segmentation module (H.S.), and head-orientation classifier (H.O.), respectively, and then used to generate the 3D location of humans, head segmentation, and head orientation, respectively. Combining this information produces an integrated description of how people are relatively positioned in space and how their heads are oriented.

This paper is organized as follows. Section 2 describes the preprocessing stages for extracting multiple features. Section 3 describes the stereo algorithm to compute 3D-depth information for location of humans. Section 4 presents the head segmentation algorithm, and section 5 the head-orientation classification algorithm. Section 6 presents experimental results. Finally, in section 7, we discuss the results and future research directions.

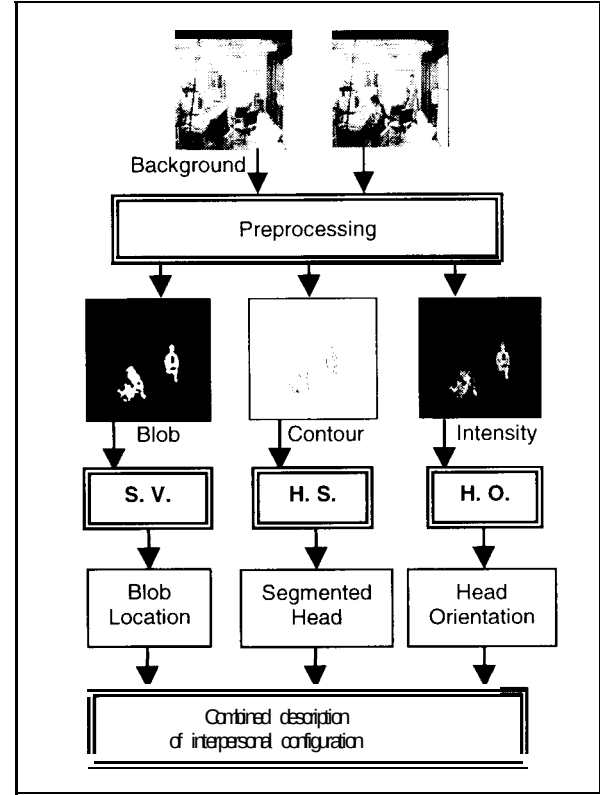


Figure 1. System Diagram (SV: Stereo vision module; HS: Head segmentation module; HO: Head orientation classifier)

2. Preprocessing for Multiple Feature Extraction

In order to use the background subtraction method for foreground / background segregation, we need a background model of the background image. Using 10 frames of background images, we calculate the mean and standard deviation of individual pixel locations for background modeling, assuming a normal distribution of intensity change in each pixel across the frames. For each image point (i, j) of the background model, the mean intensity $MB(i, j)$ and the standard deviation $SdB(i, j)$ are calculated as follows:

$$MB(i, j) = \frac{1}{N} \sum_{k=1}^N B_k(i, j) \quad (1)$$

$$SdB(i, j) = \frac{1}{\sqrt{N-1}} \sqrt{\sum_{k=1}^N (B_k(i, j) - MB(i, j))^2} \quad (2)$$

where $B_k(i,j)$ is the k -th background image intensity at pixel location (i,j) and N is the number of background images ($N = 10$).

Foreground segregation is performed at each pixel by using the background model as follows: at each image point (i,j) , the foreground pixel value $F(i,j)$ is calculated by subtracting the mean background image pixel value $MB(i,j)$ from a given input image point $I(i,j)$.

$$F(i, j) = I(i, j) - MB(i, j) \quad (3)$$

A binary image Fb is obtained by applying a multiple threshold scheme that is composed of SdB and a threshold T :

$$Fb(i, j) = 0 \text{ if } |F(i, j)| \leq 2SdB(i, j) + T \quad (4)$$

$$Fb(i, j) = 1 \text{ otherwise}$$

SdB reflects the variation of illumination in open areas where weather or natural light condition has effects. T is held constant throughout the entire processing. T plays a major role in binarization in closed areas where overall illumination does not vary. Morphological operations are applied to the binary foreground image F_b to reduce noise and produce reliable binary blobs. The ‘Canny’ edge detector is applied to the binary blobs to produce the silhouette contour image. The intensity image is obtained by masking the foreground binary blob F_b to the original input image I .

3 Stereo Vision to Locate Human Figures

The stereo vision module uses a parallel-axis stereo camera system using a pinhole camera model to produce a coarse depth map of the foreground binary blobs. Depth d of each foreground blob in an image is calculated as follows:

$$d = B \frac{f}{D} \quad (5)$$

where B is the baseline, f is the focal length, and D is the disparity.

Each person’s location is represented as a point in space by computing the depth of the foreground blob in an image frame. The depth map is computed with triangulation under the constraints of left-right order and blob-size. A projection histogram of the binary blobs is used as input to the correspondence calculation under the 1D epipolar constraint. The stereo vision module produces a relative configuration of multiple persons in space. The effective range of depth was 2 to 10 meters, within which range the multiple persons are

contained in an image frame with practical resolution of head image parts. Figure 2 shows the binary blobs, their projection histograms, and the output of the stereo vision module.

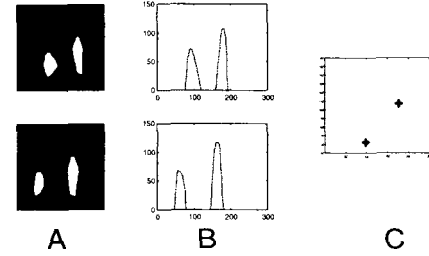


Figure 2. 3D depth of multiple persons
A: Blob information from left and right camera. B: Projection histograms. C: Depth plot (top view)

4. Head Segmentation

Head segmentation is necessary for classifying head orientation. Because our head-orientation classifier is view-based, with no specific head model, it is important to reliably crop a head portion from an input image. We propose a partial-ellipse fitting algorithm for head segmentation. Unlike normal ellipses (Equation 6 below), which are more applicable to edge maps[9], a partial ellipse is more appropriate for contour maps, where a head contour is connected to a shoulder contour. By removing the bottom portion of the normal ellipse by the amount of 30 degrees of polar angle (Equation 7), we make a region of interest (ROI) encompassing the partial ellipse and exclude shoulder pixels from the computation of ellipse fitting. Without this removal, the finally fitted ellipse is usually biased toward the neck and shoulder region of the image.

$$1 = \left(\frac{X-a}{R_x} \right)^2 + \left(\frac{Y-b}{R_y} \right)^2 \quad (6)$$

$$X = \cos(L + \pi/6), Y = \sin(L + \pi/6), \forall L \in \left[0, \frac{5}{3}\pi \right] \quad (7)$$

The ellipse fitting is performed as an optimization process to minimize the distance metric D :

$$D = \min_k \left[\sum_i^I Ds_{ik} + \sum_j^J De_{j,k} + PC_k \right], \quad \forall k \in [1, K] \quad (8)$$

where $D_{s_{ik}}$ is the distances from the i -th silhouette point to the nearest ellipse point, and $D_{e_{jk}}$ is the distance from the j -th ellipse point to the nearest silhouette point in the ROI, and PC, is the ratio of the total number of the silhouette points to the number of the fitted points when the k -th ellipse is applied. We adopt the following constraint regarding a human body: the head is located at an extreme position of the upper body image and is approximately indexed by the highest projection histogram position [5]. This constraint reduces the search space in ellipse fitting.

Figure 3 shows examples of a fitted ellipse overlaid on a foreground image. Note that the actual size of the fitted head in image C is about the half of that in image B.

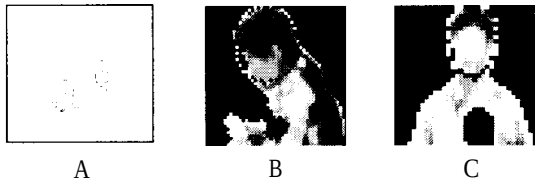


Figure 3. Head segmentation using contour
A: Silhouette contour; B, C: Fitted ellipses overlaid on foreground image.

5. Head Orientation Classification

Our head orientation classifier is view-based and does not use a specific head model or face model in order to handle an image arbitrarily oriented on a horizontal plane. Facial-feature dependent algorithms [6, X] cannot be applied to arbitrary head orientations, including the rear view and side view of the heads which lack facial features. Unlike pixel-dependent algorithms [2,3], our classifier uses moments of intensity distribution of pixels by masking a 3×4 grid on a head image (See figure 4.) The moments-based algorithm makes the classifier independent of image size and robust against pixel loss due to imperfect ellipse fitting or imperfect background subtraction. Moments are computed for each of the 12 grid elements, respectively, using nonzero pixels in the grid. Extremely bright outlier pixels due to eyeglasses or earrings etc. are excluded from the computation by using a threshold calculated based on the overall intensity distribution of the head images in the training data. The 3×4 grid generates a feature vector X composed of 12 moments values:

$$X = (W_1, W_2, \dots, W_{12}) \quad (9)$$

where each W_i is computed as follows:

$$W_i = \frac{\left(\frac{1}{N_i} \sum_{j,k} I_i(j,k) \right) - M}{SD} \quad (10)$$

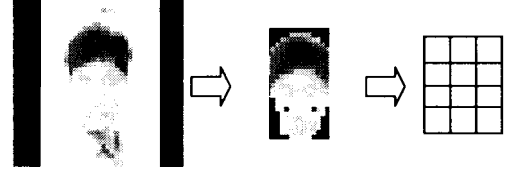


Figure 4. Moments feature extraction using a 3×4 grid after excluding outlier pixels in head image

where $I_i(j,k)$ is the pixel intensity at pixel location (j,k) in window W_i of the grid and N_i is the number of non-zero pixels in W_i . M and SD are the mean and the standard deviation, respectively, of the overall nonzero pixels in the cropped input head image.

Classification of head orientation is performed by using a K-nearest neighbor classifier [4]. We use 7 nearest neighbors out of 32 training samples for the classification. The K-nearest neighbor classifier assigns a feature vector X to class Ω , if the feature vector has the largest number of nearest neighbors belonging to class Ω among the 7 nearest neighbors. Discriminant function $g_k(X) = -d(X, X_k)$, $k \in \{1, \dots, 32\}$ is a measure of the distance between X and the feature vector X_k in the training set. The distance function $d(X, X_k)$ is selected as the L_2 norm:

$$d(X, X_k) = \|X - X_k\| = \sqrt{(X - X_k)^T (X - X_k)} \quad (11)$$

Figure 5 shows the results of classification for the segmented heads in figure 3. It shows that the classifier successfully classifies different-size heads which lack facial features and / or suffer from pixel loss.

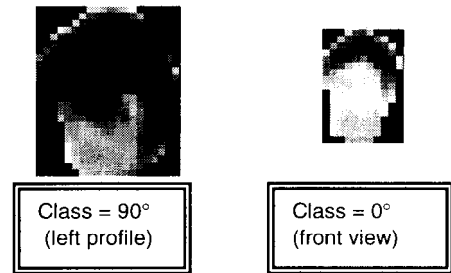


Figure 5. Examples of classification results with different-sized images.

Training Samples	Test Samples	Correct	Incorrect	%
32	64	53	11	82.8
Type of Error	Adjacent class (Error < 45°)		10	15.6
	Other classes (Error ≥ 45°)		1	1.6

Table 1. Experimental results and the type of errors

6. Experimental Results

Table 1 shows experimental success rates. 32 images were used as training data, and 64 other images as test data. Training data contains 4 images per head-orientation class, and test data 8 images per class. Table 1 shows that 82.8% of the test images were correctly interpreted. Of the 17.2% misclassified configurations, 15.6% was actually interpreted within an adjacent orientation (error $\leq 45^\circ$ of head orientation). Only 1.6% of the test data showed large error values ($45^\circ \leq \text{error} \leq 90^\circ$).

7. Conclusion

We have presented a new approach to detecting the relative configuration of multiple persons in 3D space using a stereo vision system, head segmentation method, and head orientation classifier. Our system can deal with low resolution images lacking distinguishable facial features like eyes, nose, lips, etc. The system can detect the 3D location of multiple persons, and can segment head parts in an image and classify head orientations successfully. By combining the three outputs of 3D location, head segmentation, and head orientation, we can obtain a concise representation describing how multiple persons' heads are configured in 3D space.

The combined description can be used to infer how multiple persons pay visual attention to each other in an interaction. Because any human interaction requires mutual attention, this information gives us useful prior knowledge for 'understanding' what kind of interaction occurs in the scene. It is also worth mentioning that different interaction patterns entail different visual attention patterns and, correspondingly, different head orientation patterns. Therefore, extension of the current research to consecutive changes in head location and orientation will give us more detailed information about

ongoing interaction patterns. This will be the next step of our research.

References

- [1] J. K. Aggarwal and Qin Cai. Human Motion Analysis: A review. *Computer Vision and Image Understanding*, 1998, pp. 428-440.
- [2] S. Baluja. Probabilistic Modeling for Face Orientation Discrimination: Learning from Labeled and Unlabeled Data. *Advances in Neural Information Processing Systems 11*, 1999.
- [3] T. Darrell, B. Moghaddam, and A. P. Pentland. Active face tracking and pose estimation in an interactive room, *Proc. CVPR* 1996.
- [4] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. New York Wiley, 1973.
- [5] I. Haritaoglu, D. Harwood, and L. S. Davis. Hydra: Multiple People Detection and Tracking Using Silhouettes. *Proc. Visual Surveillance*, 1999.
- [6] T. Horprasert, Y. Yacoob, and L.S. Davis. Computing 3-D Head Orientation from a Monocular Image Sequence. *IEEE Proc. Int'l Conf. Automatic Face and Gesture Recognition*. 1996.
- [7] M. S. Livingstone, and D. H. Hubel. Segregation of form, color, movement, and depth: anatomy, physiology, and perception. *Science*, 240: 740-749, 1988.
- [8] H. A. Rowley, S. Baluja, and T. Kanade. Rotation invariant neural network-based face detection. Tech. Report, Department of Computer Science, Carnegie Mellon University, CMU-CS-97-201, 1997.
- [9] S. Shirohey. Human face segmentation and identification. MS. Thesis. University of Maryland, 1993.