Absolute Head Pose Estimation From Overhead Wide-Angle Cameras

Ying-Li Tian, Lisa Brown, Jonathan Connell,
Sharat Pankanti, Arun Hampapur, Andrew Senior, Ruud Bolle
IBM T.J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532 USA
{ *yltian,lisabr,jconnell,sharat,arunh,aws,bolle* }@us.ibm.com

Abstract

Most surveillance cameras have a wide-angle field of view and are situated unobtrusively in overhead positions. For this type of application, head pose estimation is very challenging because of the limitations of the quality and resolution of the incoming data. In addition, as a person moves, their pose relative to the camera changes, but the desired pose is the absolute pose with respect to the room or space. In this paper, we present a solution to estimate absolute coarse head pose for wide-angle overhead cameras by integrating 3D head position and pose information. The work involves image-based learning, pose correction based on 3D position, and real-time multi-camera integration of low-resolution imagery. The system can be applied to an active face cataloguer to obtain the best view of the face for surveillance, to customer relationship management to record behavior in retail stores or to virtual reality as an input device.

Keywords: head pose estimation, gaze direction estimation, head tracking, face tracking.

1. Introduction

For practical applications in retail management and surveillance, overhead wide-angle cameras are necessary to view the entire scene and extract global information. Such cameras can be used to track people and their traffic patterns. Applications include improving store design, advertising effectiveness, and face cataloguing. These applications can be enhanced by knowledge of the person's activity, and one of the obvious first steps, is to understand the pose of the person.

Overhead wide-angle cameras introduce several challenges to the problem of head pose estimation. These challenges include significant lens distortion artifacts and low-resolution imagery of the head. In addition, images from such cameras introduce significant 'virtual' or relative pose. As far as we know, no prior work exists to estimate absolute coarse head pose for overhead wide-angle cameras. Figure 1 shows an individual who is looking in the same absolute direction (towards the front wall), but the head poses in the image appear differently in different positions because of the relative orientation of the wide-angle overhead camera. Notice the very low resolution of the head region (from 8x8 to 20x20 pixels) when the person is in different positions.

Figure 1. An individual is looking in the same absolute direction (towards the front wall) but appears to be looking down or to the side because of the relative orientation of the wide-angle overhead camera.

In this paper, we present a solution to estimate absolute coarse head pose for overhead wide-angle cameras by integrating 3D head position and pose information. Figure 2 shows an example of the output of this system. As a person walks around the space, their 3D location is recorded along with the absolute cardinal direction of their pose – i.e. north, east, south or west. The pose refers to which direction they are facing – that is, their head rather than their body pose.

Other researchers have investigated the use of stereo (narrow-baseline) to compute head pose [1-3], but as far as we know, no prior work exists to exploit wide baseline stereo to estimate coarse head pose for overhead wide-angle cameras. In our previous work, we performed a comparative study of methods for coarse head pose based on learning. However, in that study we evaluated two methods using a single close-up camera facing the individual [4].

In our system, an image-based head pose classifier is trained to estimate 12 pan poses from 0 to 360 degrees

using the training data from an unrelated camera. We emphasize that the training data comes from an unrelated camera so that an arbitrary camera configuration can be used. In order to obtain the absolute head pose and correct the '*virtual pose*' due to the overhead wide-angle cameras, we calculate the 3D head position based on a wide-baseline stereo method.

The 2D head positions are detected by a shape-based head finder for two calibrated overhead wide-angle cameras. The positions are used to calculate the 3D head position. Then the detected 2D head image is corrected based on camera calibration information for each camera and is fed to the general-trained head pose estimator to get the camera-based head pose. The absolute head pose for each camera can be obtained by correcting the camera-based head pose from the '*virtual pose*' which is computed based on the 3D head positions. Given an estimate of absolute pose from each camera and the 3D position of the head, a maximum likelihood estimation is performed to improve the final result.

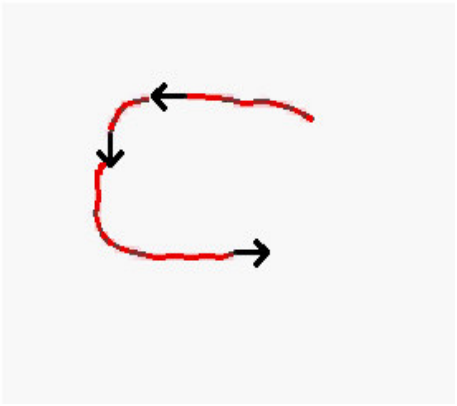


Figure 2. The system tracks the 3D position of a person as they walk around a room and their absolute cardinal head pose.

This paper is organized as following. Section 2 describes the system setup and the method to detect the 2D and 3D head positions. Section 3 discusses the '*virtual pose*' correction for overhead wide-angle cameras. Section 4 describes the method for absolute head pose estimation. This is followed by the experimental results in Section 5; finally we summarize our paper and present future directions in Section 6.

2. Head Region and Position

2.1 System Setup

Our group has built a 3D people tracking system, which is used to drive a face cataloguer and position-based video retrieval. The system is based on wide-baseline stereo control of multiple active cameras. It enables the *continuity of identity*. We associate each path with close-up images of the particular person, allowing us to answer: "who is where?" at any point of time [5,6]. Figure 3 shows the positions (in mm) of the two cameras and their orientations and the path a person took as they walked around the room (center). The coordinate system of camera 1 is shown at right, the coordinate system of camera 2 is shown in the upper left corner.

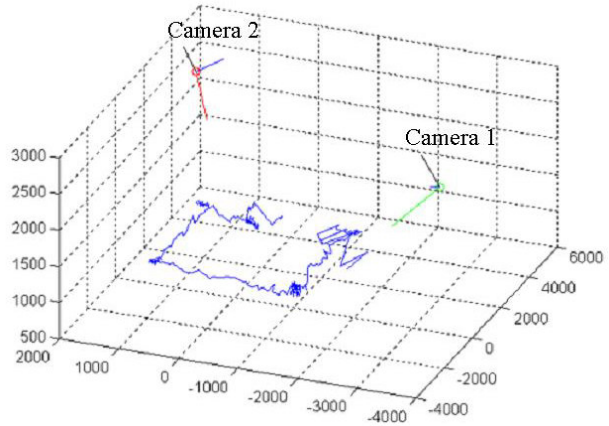


Figure 3. The track of a person walking around the room and the positions (in *mm*) of the two cameras and their optic axes.

2.2 Background Subtraction

As shown in Figure 4, the background subtraction module combines evidence from differences in color (top left), texture (middle left), and motion (bottom left). The use of multiple modalities improves the detection of objects in cluttered environments. The resulting saliency map (top right) is smoothed using morphology-like operators and then small holes and blobs are eliminated to generate a clean foreground mask (middle right).

The background subtraction module has a number of mechanisms to handle changing ambient conditions and scene composition. First, it continually updates its overall RGB channel noise parameters to compensate for changing light levels. Second, it estimates and corrects

for AGC and AWB shifts induced by the camera. Finally, it maintains a map of high activity regions (lower right) and slowly updates its background model only in areas deemed as relatively quiescent.

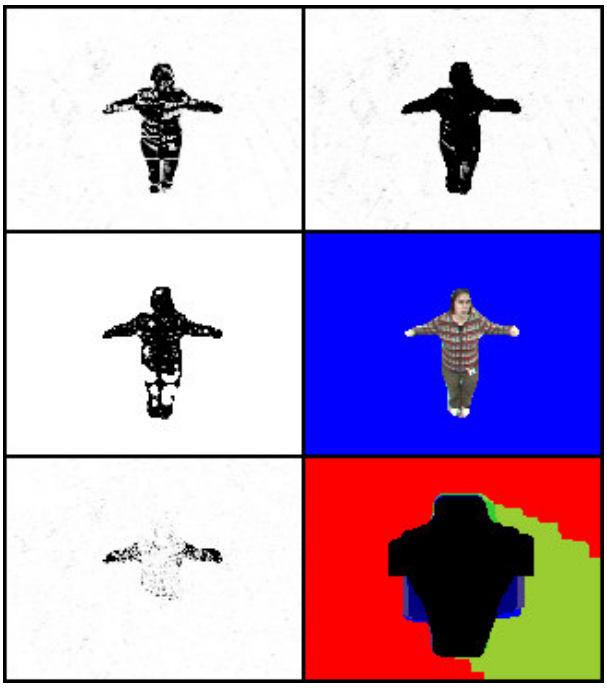


Figure 4. Background subtraction combines evidence from color, texture and motion.

2.3 Head Localization and 3D Head Position Detection

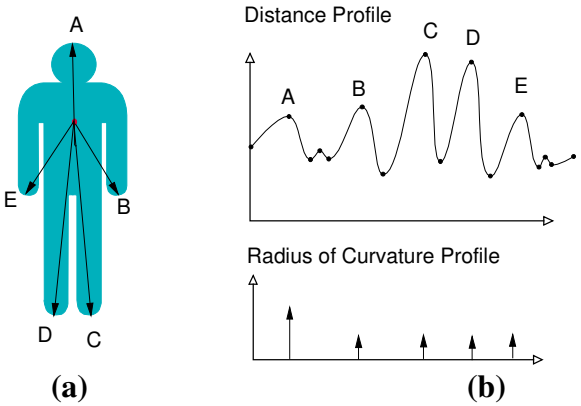


Figure 5. Head detection steps. (a) The silhouette information (b) Distance profile showing significant peaks and the radii of curvature at the significant peaks.

The head detection uses the smoothed silhouette of the foreground object as segmented using background subtraction. To interpret the silhouette, we use a simple human body model consisting of six body parts: head, abdomen, two hands, and two feet as shown in Figure 5. First, we generate a one-dimensional "distance profile" that is the distance of each contour pixel from the contour centroid, following the contour clockwise. This distance profile is parsed into peaks and valleys based on the relative magnitudes of the successive extreme. The peaks of the distance transform are used to hypothesize candidate locations of the five body parts: the head, two feet, and two hands. Determination of the head among the candidate locations is currently based a number of heuristics based on the relative positions of the candidate locations and the curvatures of the contour at the candidate locations. More specifically, the following objective function is used to decide the location of the head:

$$O_i = (Y_i - Y_c) + w_x * |(X_c - X_i)| + w_r * R_i - w_e * E_i,$$

where (X_c, Y_c), (X_i, Y_i) denote the co-ordinates of the centroid of the body contour and center of the circle fitted to the contour segment associated with i^{th} peak. R_i, E_i denote radius and residue of least square fitting of the i^{th} circle. $w_x(=1)$, $w_r(=1)$, and $w_e(=10)$ are weights associated with the three components of the objective function. In other words, the objective function hypothesizes that smaller, more circular extreme are more likely to be heads. Similarly, the circles that are higher and vertically more aligned with the center of the body are preferred as heads. Our approach is similar to [7].

Based on the 2D head position on each camera and the camera calibration information, the 3D head position is derived using a wide baseline stereo method.

3. Virtual Pose Correction for Overhead Wide-angle Camera

The head pose estimation for overhead wide-angle cameras needs to correct for '*virtual pose*' using the 3D head and camera positional information. In our approach, the imagery from two cameras and their relative geometry are used to improve performance.

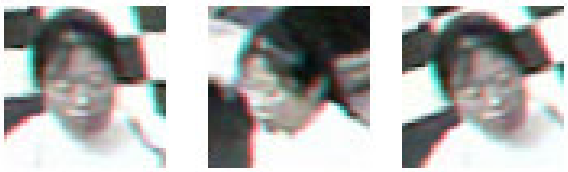


Figure 6. Virtual tilt (looking up/down), virtual pan (looking left/right) and virtual roll (image rotation) for overhead wide-angle camera.

We use the terms: tilt, pan, and roll to refer to the head looking up/down (tilt), left/right (pan), and the rotation

along the optic axis (roll). Figure 6 shows an instance of *virtual tilt* (left image), *virtual pan* (center image), and *virtual roll* (right image).

3.1. *Virtual Tilt* Correction

Prior to designing the algorithms for *virtual tilt* correction, we computed the sensitivity of *virtual tilt* with regards to relative position of the person with respect to the camera. Figures 7 shows the results of this analysis. In this figure, we show how *virtual tilt* is determined by the distance to the camera depending on the height of the person and the height of the camera. As a person gets very close to the camera, virtual tilt becomes extreme. On the other hand, for indoor cameras, in rooms of typical heights (8-9 feet) and people of normative size, virtual tilt can be kept below 20° for distances greater than approx. 6 feet from the camera or below 10° for distances greater than 12 feet.

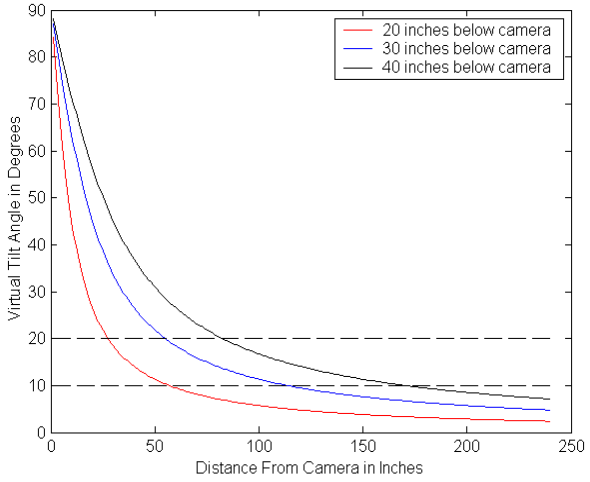


Figure 7. The sensitivity of *virtual tilt* regards to relative position (in *inches*) of the person with respect to the camera.

Virtual tilt is computed based on the camera origin $\vec{c}$ and the position of the person's head $\vec{h}$. From these positions, we compute the epipolar line: $\vec{e} = \vec{h} - \vec{c}$. We assume the cardinal directions (or the directions of interest) align with those of the x and y axes of the identity matrix in the world coordinate system. Virtual tilt θ_x is the angle between the epipolar line and the z-axis:

$$\theta_x = \cos^{-1}\left(\frac{\vec{e} \bullet [0,0,1]^t}{\|\vec{e}\|}\right)$$

Note, since virtual tilt refers to tipping the head up or down, it is with respect to some direction of the face. Hence, if a person is turned with respect to the camera , virtual tilt is a tip of the head in this direction. In our system**,** we train only on a fixed tilt ($\theta_x = 10°$) with respect to the frontal face.

3.2. *Virtual Pan* Correction

Figure 8 shows the sensitivity of *virtual pan* with regards to relative position of the person with respect to the camera. *Virtual pan* depends on how close to the camera (the horizontal axis in Figure 8) and how far to the left (or right) of the camera (the axis into the page in Figure 8). Again, the sensitivity is most extreme when a person is near the camera but is nearly linear beyond a few feet.

Virtual pan θ_y can be computed similarly to *virtual tilt* using the complement of the angle between the epipolar line and the y-axis of the world coordinate system.**:**

$$\theta_y = \cos^{-1}\left(\frac{\vec{e} \bullet [0,1,0]^t}{\|\vec{e}\|}\right).$$

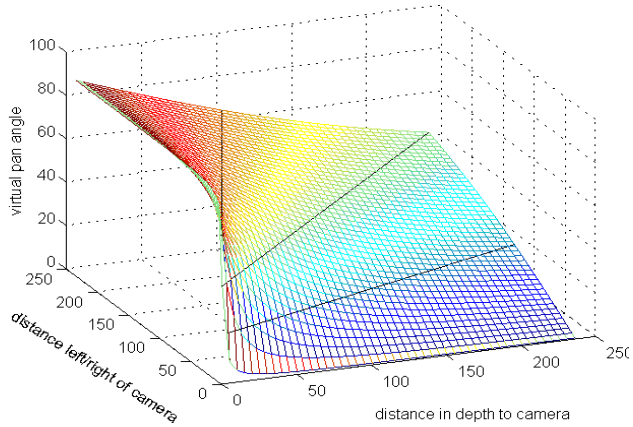


Figure 8. Sensitivity to *virtual pan* regards to the relative position (in *inches*) of the person with respect to the camera.

3.3. *Virtual Roll* Correction

Virtual roll θ_z, is computed based on the angle between the vertical axis of world coordinate system projected onto the camera's image plane and the vertical axis of the camera system:

$$\theta_z = \cos^{-1}\left(\frac{\left(R * [0,0,1]^t + T\right) \bullet [0,1,0]^t}{\left\|R * [0,0,1]^t + T\right\|}\right)$$

where (R,T) represent the rigid transformation from the world coordinate system to the camera coordinate system.

4. Absolute Head Pose Estimation for Overhead Wide-Angle Cameras

4.1 Absolute Head Pose Estimation

The flow chart of the algorithm used to estimate absolute pose by the general camera configuration system is shown in Figure 9. Each overhead wide-angle camera is independently calibrated. Based on the calibration information, each camera corrects for virtual roll using image rotation. Then the system uses background subtraction and the head finder to find the head region in each image. Relative pose estimation is then performed by a neural network based head pose classifier for each camera [4]. We call this 'relative' pose estimation, because at this point, the algorithm estimates pose based on appearance only for each camera.

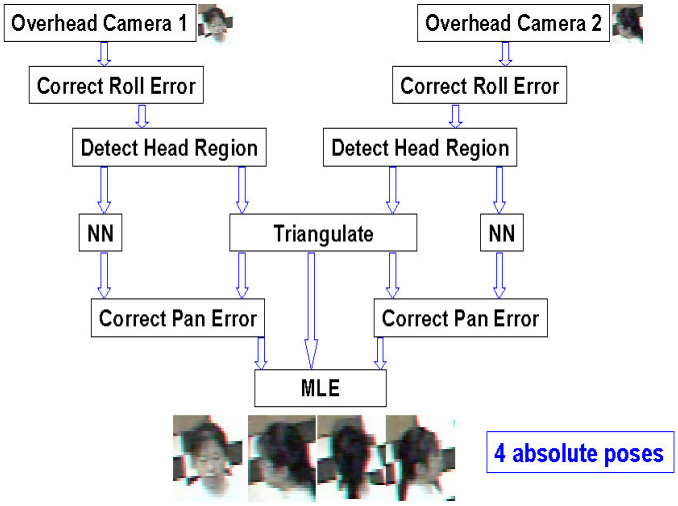


Figure 9. Overview of the method to estimate absolute head pose for overhead wide-angle camera.

Based on the camera calibration information and the 3D head position, the system can now correct for *virtual pan* to obtain an estimate of absolute pose. Given an estimate of absolute pose from each camera and the 3D position of the head, maximum likelihood estimation is performed to obtain the final absolute head pose. The maximum likelihood estimation assumes the two views of the head are seen from two cameras with known geometry, and thus the angle between the two views is known.

The final estimate of pose maximizes the likelihood that both cameras are viewing this pose from their respective vantage points. The probability distribution $P(Z_k \mid \theta)$ of an image observation Z_k for the k^{th} camera given a specific pose θ, depends on camera noise, lighting variations, individual variations, etc. The majority (although not all) of these factors are independent for each camera. Therefore, we model the joint probability as the product of the individual distributions:

$$P(Z_1, Z_2 ... Z_n \mid \theta) \approx \prod_k^n P(Z_k \mid \theta).$$

The outputs of each individual pose estimator are inversely weighted by the distance of the head from that camera. In this way, the system can bias the output based on the closer camera.

4.2 Non-Uniform Head Pose Ground Truth Data

In the generalized camera configuration system, we acquire training data from an independent wide-angle overhead camera. This data was acquired at the following pan angles in degrees:

(0,15,30,49,90,120,180,-120,-90,-49,-30,-15).

Our prior system was based on the CMU Pose, Illumination and Expression database [ref] which is in uniform angles. An example of non-uniform angle sampling (from 0-90°) is shown in the bottom row of Figure 10.

Figure 10. Top row: uniform angle sampling (0°,22.5°,45°,67.5°,90°). Bottom row: non-uniform sampling (0°,15°,30°,49°,90°).

There are two reasons that training data was acquired at non-uniform angle spacing. First, since images of the face are projected onto the image plane, the spacing between the features of the face does not change linearly with pan angle. In particular, we use pan angles, which cause the projection of the nose (center of the face) to move in equal step sizes from the center to the outer tangent of the face (i.e. when the face is turned 90°). To visually contrast the two sampling techniques, uniform sampling is shown in the top row of Figure 10. Notice the similarity between the views at 67.5° and 90°.

Secondly, as the subject turns from 90° to 180° we train additionally only at 120°. If you assume that half the head is covered by hair (the back 180° hemisphere) then half the face can be seen when turned to 90°.

Similarly, a quarter of the face can be seen at 120°. This non-uniform sampling of pan angle space is important in order to optimize the ability of the classifier to discern pose.

We performed experiments based on two sets of data. The initial database used was the CMU Pose, Illumination and Expression database [14] composed of 68 subjects. This dataset contains high resolution close-up images. In our previous work, we discuss some of the issues regarding the effectiveness of the dataset to generalize to a larger population and how these issues can be addressed [4]. We believe that training data needs to be acquired for a wide range of prototypes and ultimate systems will need to differentiate (implicitly or explicitly) prototypes in order to correctly classify pose given the enormous range of variation in personal characteristics.

On the other hand, once an appropriate prototype match is found, classification can proceed from the smaller but more relevant database. This tactic was utilized in our wide-angle overhead experiments. In these experiments only data from a small number of subjects acquired in our laboratory was used.

5. Experimental Results

Results of our first experiment were based on the CMU PIE database and are shown in Table 1. In this experiment, images were classified into 9 uniformly-spaced poses ranging from 0-180°. The "two camera angle," i.e., the angle from one camera to the head to the other camera, is known. Since the subject is centered in the view of each camera, no virtual pose is evident. This simple experiment was used to verify our assumption that classification could be improved using multiple cameras. On average, performance improved from 89% correct classification to over 96% for a single image at 32x32 to from 79% to 85% for a single image at 16x16.

Two camera angle (degrees)	Head Resolution	
	32x32	16x16
22.5	96%	85%
45	97%	85%
67.5	98%	90%
90	95%	80%
Single camera	89%	79%

Table 1. Result of pose classification for two cameras using CMU PIE database

The remaining experiments were performed with data obtained in our laboratory. In these experiment overhead wide-angle cameras are used. Pose classification follows the algorithm outlined in Section 4. Table 2 shows a summary of the absolute head pose classification results based on 4799 images.

		Cam1	Cam2	Both
Recognition Rate	No Correction	76.4%	65.6%	81.3%
	Virtual Pose Correction	80.6%	69.1%	84.3%
Average Head Size (pixels)		18x18	13x14	
Average Distance between Head and Camera (M)		3.65	4.48	

Table 2. Results of absolute head pose estimation

In our results, the average head size is 18x18 pixels for Camera 1 and 13x14 pixels for Camera 2. The average distance between the head and the Camera 1 and Camera 2 is 3.65m and 4.48m respectively. In Table 2, the columns "*Cam1*" and "*Cam2*" list the outputs of each individual head pose estimator. The final head pose results using the maximum likelihood estimator combining the individual outputs from both cameras are listed in the column "*Both*". Without "*virtual pose*" correction, an average recognition rate of 76.4% was achieved for camera 1 and a recognition rate of 65.6% for camera 2. By combining the outputs of camera 1 and camera 2, the recognition rate increased to 81.3%. With "*virtual pose*" correction, the average recognition rates of the camera 1, camera 2 and the combination of both cameras increased to 80.6%, 69.1%, and 84.3% respectively.

More examples of our algorithm applied to image sequences taken in our lab can be viewed at http://www.research.ibm.com/peoplevision. In Figure 11 we show three examples. In each example, the top left image is from Camera 1, the top right image is from Camera 2, and the center white square shows the path of the person in red (in 2D) ending with the estimated absolute head pose (black arrow). Currently the absolute head pose is identical to the head pose from Camera 1's point of view. The gauge at left shows in red, the relative pose for Camera 1 (from Camera 1's point of view), in green, the virtual pose, and in blue the absolute pose. The virtual pose represents the amount of pose change expected due to the position of the person with respect to the camera. The gauge at right shows the analogous estimates for Camera 2 (from Camera 2's point of view). The bottom images show the detected head regions from each camera enlarged for visualization.

The three examples illustrate typical system behavior. In Figure 11(a) both individual cameras correctly estimate pose. Notice the correction from relative (red) to absolute pose (blue) based on the virtual pose (green). When the subject is equidistant to both cameras,

integrated pose estimation is driven by both distributions (Figure 11(b)). When the subject is near one camera, its peak probability dominates the integrated pose estimation(Figure 11(c)).

Figure 12 shows (from left to right) a spatial contour plot of accuracy for Camera 1, Camera 2, and the integrated maximum likelihood estimate. The positions of the cameras are shown as diamonds at top (Camera 2) and at right (Camera 1) in each figure. For Camera 1, the accuracy falls off near the position of Camera 2 (diamond at top). For Camera 2 the accuracy decreases as the subject is more distant from the camera. The integrated result clearly shows the advantage of combining pose information in 3D. It is interesting to note, that in general, the accuracy falls off sharply.

6. Conclusions

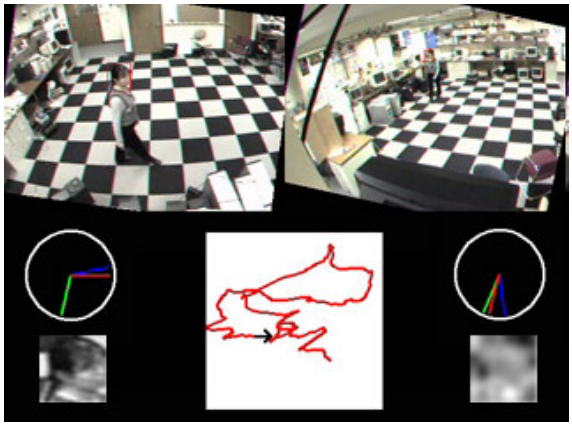
In this paper, we successfully integrated head pose estimation for overhead wide-angle cameras with 3D position information. Unlike previous work on head pose estimation, the overhead wide-angle cameras provide very low-resolution imagery of the head, introduce significant 'virtual' or relative pose and significant lens distortion artifacts. To enable arbitrary camera configurations, the head pose estimator was trained on an independent camera configuration. In order to obtain correct absolute head pose for overhead wide-angle camera, the virtual pose is corrected based on the head position and pose information. An average recognition rate of 85% was achieved.

References

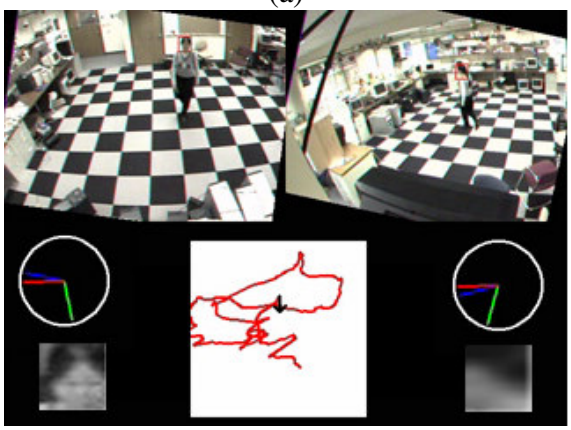
[1] Y. Matsumoto and A. Zelinsky, An Algorithm for Real-time Stereo Vision Implementation of Head Pose and Gaze Direction Measurement, in *Proc. of the Int'l Conf. on Automatic Face and Gesture Recognition*. Grenoble, France, March 28-30, 2000, Pages:499-504.

[2] Ming Xu and T. Akatsuka, Detecting head pose from stereo image sequence for active face recognition, in *Proc. of the Int'l Conf. on Automatic Face and Gesture Recognition,* 14-16 Apr 1998,Page(s): 82 –87.

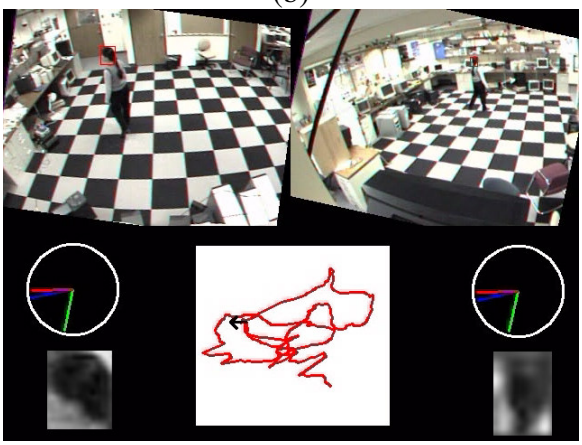
[3] R.Yang and Z Zhang, Model-based head pose tracking with stereovision, in *Proc. of the Int'l Conf. on Automatic Face and Gesture Recognition,* 20-21 May, 2002, Page(s): 242 –247.



(a)



(b)



(c)

Figure 11. Three Examples of absolute head pose estimation

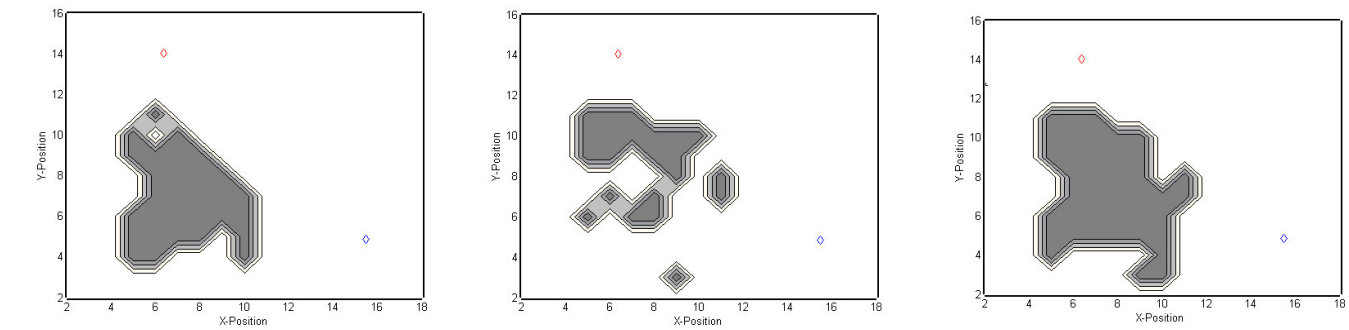


Figure 12. Spatial contour plots of accuracy for Camera 1, Camera 2, and the integrated system.

[4] L. Brown and Y-L Tian, "Comparative Study of Coarse Head Pose Estimation," IEEE Workshop on Motion and Video Computing, Orlando FL, December 5-6, 2002, pp125-130.

[5] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, and R. Bolle, "Face Cataloger: Multi-scale Imaging for Relating Identity to Location," IEEE Conf on Advanced Video Surveillance and Signal Based Surveillance, Miami 2003.

[6] A. Senior, A. Hampapur, Y.-L. Tian, L. Brown, S. Pankanti, and R. Bolle, "Appearance Models for Occlusion Handling," in *Proc. 2nd IEEE Int'l Workshop on Performance Evaluation of Tracking in Surveillance*. Kauai, HI, Dec 9, 2001.

[7] H. Fujiyoshi and A. Lipton, Real-time Human Motion Analysis by Image Skeletonization, Proc. of the Workshop on Application of Computer Vision, October, 1998.

[8] S. Basu, I. Essa, and A. Pentland, "Motion Regularization for Model-Based Head Tracking," in *13th Int'l Conf on Pattern Recognition*. Austria, Vienna, August 25-30, 1996.

[9] J. Heinzmann and A. Zelinsky, "3-D Facial Pose and Gaze Point Estimation Using a Robust Real-Time Tracking Paradigm," in *Proc. of the 3rd Int'l Conf. on Automatic Face and Gesture Recognition*. Los Alamitos, CA, April 14-16, 1998, pp. 142-7.

[10] T. Jebara and A. Pentland, "Parameterized Structure from Motion for 3D Adaptive Feedback Tracking of Faces," in *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[11] M. La Cascia, S. Sclaroff, and V. Athitsos, "Fast, Reliable Head Tracking Under Varying Illumination: An Approach Based on Registration of Texture-Mapped 3D Models," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, April, 2000.

[12] M. Malciu and F. Preteux, "A Robust Model-Based Approach for 3D Head Tracking in Video Sequences," in *Proc. of the Fourth Int'l Conf. on Automatic Face and Gesture Recognition*. Grenoble, France, March 28-30, 2000, pp. 169-74.

[13] Z. Zivkovic and F. van der Heijden, "A Stabilized Adaptive Appearance Changes Model for 3D Head Tracking," in *Proc of the 2nd Int'l Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Realtime Systems*. Vancouver, Canada: IEEE, July 13, 2001, pp. 175-181.

[14] T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) Database of Human Faces," in *Proc. Int'l Conf. on Automatic Face and Gesture Recognition*. Washington, DC, May 20-21, 2002.