CALIBRATING HEAD POSE ESTIMATION IN VIDEOS FOR MEETING ROOM EVENT ANALYSIS

Jilin Tu, Thomas Huang
Beckman Institute
Univ. of Illinois at Urbana-Champaign
405 North Mathews Avenue
Urbana, IL 61801 USA

Yingen Xiong, Travis Rose, Francis Quek
Department of Computer Science
Virginia Poly. Inst. and State Univ.
621 McBryde Hall, MC 0106
Blacksburg, VA 24061

ABSTRACT
In this paper, we study the calibration of head pose estimation in stereo camera setting for meeting room video event analysis. Head pose information infers the direction of attention of the subjects in video, therefore is valuable for video event analysis/indexing, especially in meeting room scenario. We are developing a multi-modal meeting room data analyzing system for studying meeting room interaction dynamics, in which head pose estimation is one of the key components. As each subject in the meeting room can be observed by a pair of stereo cameras, we do 2D head tracking for the subject in each camera, and the 3D coordinate of the head can be obtained by triangulation. The 3D head pose is estimated in one of the camera coordinate system, we develop a procedure to accurately convert the estimated 3D pose in the camera coordinate system to that in the world coordinate system. In the experiment, visualization of the estimated head pose and location in world coordinate system verifies the soundness of our design. The estimated head pose and 3D location of the subjects in the meeting room allows further analysis of meeting room interaction dynamics, such as F-formation, Floor-control[1], etc.

Index Terms— Video signal processing, Stereo vision, Machine Vision

1. INTRODUCTION
As the behaviors of the subjects are typically predictable and the changes of the appearances of the people and the background are relatively consistent for modeling with the state-of-the-art statistical vision techniques, researchers are gaining interests in developing systems for meeting analysis with audio and video cues. As meeting participants employ various multi-modal channels(i.e., speech, gaze, gesture, posture, etc.) to communicate with each other, the inference of these information reveals static and dynamic social structuring of the meeting participants, the flow of topics being discussed, and social dynamic units among the participants [2]. Such functionality is particularly valuable to applications, such as meeting video database indexing / management [3], psycholinguistic study of multiparty discourse dynamics [1]. For this purpose, several multi-modal meeting room corpora have been collected [4][2] and research has been conducted on automatic transcription of meetings [5], focus of attention analysis [6], action tracking [4], etc.

We are developing a planning meeting video analysis system for the purpose of psycholinguistic study[2]. The setup of our meeting room is shown in Figure 1. In the meeting room, up to 8 participants, A, B, C, D, E, F, G, and H, sit around a rectangular conference table. 10 Canon GL2 camcorders(C1-10) surrounding the room are mounted in the ceiling so that each participant can be captured by at least two camcorders. Calibration is carried out between 12 camcorder pairs and between the camera and the room world coordinate system so that 3D information of each participant can be obtained. Typical views of the participant E from camcorder C1 and C3 are shown in Figure 2.

Given the setup, we do 2D head tracking for the subject in each camera, and the 3D coordinate of the head can be obtained by triangulation. The 3D head pose is estimated in one of the camera in which the subject assumes near-frontal

Fig. 1. The camera setup for the meeting room
Fig. 2. Views of subject E from camcorder pair C1 and C3

view. The estimated 3D pose is then converted into that in world coordinate system. After doing this to each subject, the calibrated pose information in the world coordinate system allows user to do further analysis on meeting room interaction dynamics, such as F-formation, Floor-control[1], etc.

2. TRACKING 3D HEAD POSITION AND HEAD POSE IN WORLD COORDINATE SYSTEM

2.1. Problem statement

Given a world coordinate system \( \Pi_w \) in the meeting room and 2 cameras with coordinate systems \( \Pi_i^c, i = 1, 2 \). We wish to estimate the 3D location \( \hat{X}_w = (X_w, Y_w, Z_w) \) and orientation \( \hat{\Phi}_w = (\theta_w, \phi_w, \psi_w) \) of the subject’s head in the world coordinate system given video observations of the two cameras.

2.2. The design of the system

The 3D location of the head \( \hat{X}_w \) can be obtained by stereo triangulation of the 2D head centroid \((x_i^u, y_i^u)\), \(i = 1, 2\). The head centroid can be estimated respectively in each camera by 2D head tracker. Several factors have to be taken into account for the estimation of 3D head pose \( \hat{\Phi}_w \). First of all, a 3D face tracker is able to estimate the head pose variations (denoted as \( \Phi_t \)) with respect to neutral head pose. The neutral head pose \( \Phi_h \) should then be estimated according to the 2D head position \((x_u, y_u)\) in the camera coordinate system. At last, the relative pose has to be converted into absolute head pose in world coordinate system. The design is illustrated by Figure 3.

![Fig. 3. The design of our system](image)

2.3. Estimating 3D head location \( \hat{X}_w \)

We track the 2D head location at each frame using a 2D tracking algorithm with incremental subspace learning[7] after the face location is provided in the first frame. It was shown this algorithm is capable to track object of varying appearances by incrementally updating its \( L_\infty \)-norm subspace model based on the past observations. The 3D head location \( \hat{X}_w \) can be estimated by stereo triangulation.

2.4. Estimating 3D head pose \( \hat{\Phi}_w \)

2.4.1. Tracking the head pose in single view camera

In our meeting room setup, the near frontal view face of the meeting participant will always be captured by one of the cameras (namely the primary camera). We chose to estimate the head pose from the primary camera. Due to the difficulties to find accurate correspondence on human face/head for the stereo camera as face/head is usually smooth and lack of texture, estimating head pose by stereo vision is a daunting task. With a prior knowledge about human head geometry and texture, estimating head pose from single view camera has advantage over estimating the pose from stereo camera and has become the dominant approach in literatures.

To do 3D model based head pose tracking from single view camera, some prominent 2D facial features are first located by Active Shape Model(ASM) in the frame when the near-frontal face of the subject is detected. A generic 3D face model is then customized in scale and orientation to fit to the face. After the initialization, various tracking algorithms can be employed to track the head pose [8][9][10].

It is typical that human faces have very low resolution in meeting room videos. Optical flow based tracking approach no longer works in such low resolution video. In [11], we proposed to estimate the 3D head poses by modeling the shape-free facial textures acquired from the video with subspace learning techniques. As shown in Fig. 4, a generic 3D face model is utilized to acquire the shape free texture from the video frame after the model is aligned to the face. The uncertainty of texture acquisition at each vertex on the face surface is modeled by surface normal direction. We then propose to model the facial appearance variations online by incremental weighted PCA subspace with forgetting mechanism, and we do the tracking in an annealed particle filtering framework. Figure 5 shows some key frames of the tracking result for a NIST video.

2.4.2. Converting the head pose in camera coordinate system into that in world coordinate system

Define the neutral head pose of the subject as a upright frontal view of the face in the primary camera, the 3D head pose tracker estimates head pose variations with the neutral head pose as origin, i.e. \( \Phi_t^{\text{neutral}} = (0, 0, 0) \). As we described in
the system design, the head pose in world coordinate system can be obtained after the neutral pose in the camera coordinate system is estimated and the conversion from camera coordinate system to world coordinate system is provided.

As shown in Figure 6, the orientation of the neutral pose in the camera coordinate system can be estimated as a function of the location of the face \((x_u, y_u)\) in the video as follows:

\[
\begin{align*}
\phi^b_x & = \arctan \frac{y_u}{f} \\
\phi^b_y & = -\arctan \frac{x_u}{\sqrt{f^2 + y_u^2}} \\
\phi^b_z & = 0
\end{align*}
\]

where \(f\) is the focus length.

Denote the rotation matrix for tracked head pose variations \(R(\Phi_t)\), that for the neutral pose as \(R_n = R(\phi^h_x, \phi^h_y, \phi^h_z)\), and the camera rotation matrix in the world coordinate system as \(R_c\), the rotation matrix for the head pose in the world coordinate system is

\[
R(\hat{\Phi}_w) = R(\Phi_t)R(\phi^h_x, \phi^h_y, \phi^h_z)R_c
\]  \hspace{1cm} (1)

3. EXPERIMENTS

We applied our algorithms to obtain the head pose and location of a subject as participant G in the meeting room with the stereo camera pair C2 and C5. In the video, the subject first placed a calibration box in front of him and then turned and moved away. In order to verify the estimation accuracy, we reconstructed the calibration box, together with an edge of the desk and a rectangular patch on a tool chest behind the subject. Two views from the 2 cameras(terget with tracking result) is shown in Figure 7. In Fig. 7-(a), the green outline marks the 3D structure of the scene we reconstruct for the visualization. In Fig. 7-(b), the view actually is enlarged for tracking 3D head poses. The white boxes indicate 2D head location tracking windows. The yellow 3D face mask with red arrow indicate the 3D head pose estimated from Camera C2. Figure 8 shows the several key frame of the 3D reconstruction of the scene, the location and direction of the 2 cameras, and that of the subject’s head. The cross X indicate the location of the cameras(head), and the triangle \(\triangle\) indicate the direction of the camera(head). And the corresponding camera view at camera C2 is also visualized for comparison. While we don’t have ground truth data for numerical evaluation yet, the visualization give us a subjective evaluation of the soundness of our proposed design. Figure 9 shows some pose estimation results for a real meeting scenario. The visualization of the scene is drawn with near top-down view angle. The human heads are indicated by spheres. The long vector indicates the head pose of the meeting host. And the shorter vectors shows the camera pose in the scene.
Fig. 7. Two views of the stereo camera (together with tracking results).

Fig. 8. The visualization of the sequence in the scene with camera view from C2. The cross X indicate the location of the cameras (head), and the triangle △ indicate the direction of the camera (head).

4. CONCLUSIONS

In this paper, we presented our meeting room video collection and analysis project. In particular, we demonstrated how we obtain the 3D head location and pose in the world coordinate system from stereo camera. The experiment result verifies the soundness of our design. We will further apply this tool to do automatic video analysis for psycholinguistic study of meeting room social dynamics[1].

Acknowledgement

This research has been supported by the Advanced Research and Development Activity ARDA VACEII grant 665661: From Video to Information: Cross-Model Analysis of Planning Meetings.

5. REFERENCES


