Head Orientation Modeling:

Geometric Head Pose Estimation using Monocular Camera

Muhammad Sikandar Lal Khan^a, Shafiq Ur Rehman^a, Zhihan LV^a, Haibo Li^b

^aImmersive Interaction Lab (i2Lab), Umeå University, Umeå, 90734, Sweden

^bSCSC, Royal Institute of Technology (KTH), Stockholm, Sweden

Email: muhammad.sikandar.lal.khan, shafiq.urrehman, zhihan.lv @umu.se

Abstract

In this paper we propose a simple and novel method for head pose estimation using 3D geometric modeling. Our algorithm initially employs Haar-like features to detect face and facial features area (more precisely eyes). For robust tracking of these regions; it also uses Tracking- Learning- Detection(TLD) frame work in a given video sequence. Based on two human eye-areas, we model a pivot point using distance measure devised by anthropometric statistic and MPEG-4 coding scheme. This simple geometrical approach relies on human facial feature structure on the camera-view plane to estimate yaw, pitch and roll of the human head. The accuracy and effectiveness of our proposed method is reported on live video sequence considering head mounted inertial measurement unit (IMU).

Keywords: Head pose estimation; 3D geometric modeling; human motion analysis.

1. Introduction

Head pose estimation is vital in many head gesture interfaces and machine vision applications such as face modeling [1], gaze estimation [2, 3] and affect analysis [4]. Most of these applications require accurate 3D orientation of human head. 3D orientation of human head using monocular camera (i.e. 2D information with absence of depth) is always a challenging task [5]. Automatic analysis and synthesis of head movement from 2D image(s) is an active field and for practical reasons researchers assume that head movement can be modeled as a disembodied rigid object with 3 degree of freedom (DOF) i.e., yaw, pitch and roll [5]. In general, human head pose estimation approaches can be broadly categorized in eight classes namely; Appearance-template, detector-array, nonlinear-regression,

manifold-embedding, tracking, hybrid, geometric methods and finally flexible models (see [5] for more details).

Appearance-template based approaches, also called key frame based approaches, use information previously acquired about the user (automatically or manually) to estimate the head position and orientation [12]. Detector-array approaches calculate the pose of head by training a series of head detectors using SVM, Adaboost cascades and neural networks [13,14]. Nonlinear-regression approaches build a relation of face images with head poses from a large number of training data [15]. Manifold embedding approaches also called dimensionality reduction based approaches, where linear/non-linear embedding of the face images is used for pose estimation [16,17]. Tracking based approaches also called motion based approaches, track the position and orientation of the head through video sequences using pair-wise registration [18]. Flexible models are those that fit non-rigid model to the facial features of each individual in the image plane for example active shape model [19], active appearance model [20] and CLM [21].

Geometric methods consider facial feature (such as eyes corners, nose tip, and lip corners) locations and their relative orientation-relationship to estimate the head pose [22, 23, 24, 25]. As these methods depend on accurate detection of facial landmarks, hence all most all geometric methods have difficulty to cope with occlusion. Our approach significantly differs from previously proposed geometric methods as it considers relationship between two eyes area and whole frontal face area in order to calculate pivot point as origin of head movement. It also employs Tracking Learning Detection (TLD) approach to successfully overcome occlusion problem.

In this paper, we propose a novel and simple approach for pose estimation using a monocular camera by exploiting eyes region relationship with human face area to estimate the head pose. We estimate the distance between the eyes

and use facial animation parameters provided by MPEG-4 [6] to estimate a pivot point (i.e., 3D origin of face). The head motion is made possible by movement of the skull with cervical vertebrae known as C2 acts as the pivot point for the motion [7]. Once the origin of human head is estimated we used our geometric based approach to extract yaw, pitch and roll angles. Section II describes our head modeling approach. Section III describes the results of experiments. Finally in section IV we present analysis of our algorithm with concluding remarks.

2. Head Orientation Modeling

Our approach proposes a computational model for head-pose based on anthropometric model and facial geometry using MPEG-4 facial animation parameters [6, 7]. Our algorithm models a coordinate system based on calculated pivot point being fixed at the cervical vertebrae. Formally, the image plane coincides with the camera's XY-plane and the viewing direction coincides with the Z-axis, where α , β and γ (yaw, pitch and roll, respectively) denote the three rotation angles of human head about the Y, X and Z axis, respectively. Here O_{yaw} and O_{pitch} denote the pivot point of yaw and pitch, and E_{Left} , E_{Right} and E_{Mid} denote left-eye center, right-eye center and eyes midpoint, respectively. The coordinates of each point be $\{X, Y, Z\}$ (see Fig. 1). Our framework consists of three steps. Firstly, eyes and face regions are estimated from monocular 2D image. Secondly, we calculate 3D pivot points for yaw and pitch based on anthropometric statistic and MPEG-4 facial parameters. Finally, to estimate the head orientation we employ TLD approach to track the detected facial feature regions and their corresponding points of interests in successive frames.

Fig. 1. From left to right - The three degrees-of-freedom framework used for representation of head orientation in current work (i.e. Yaw, Pitch and Roll); and estimated facial points (E_{Left} , E_{Mid} , E_{Right}) and pivot points (O_{yaw} , O_{pitch}) detected for tracking in successive frames using TLD.

2.1 ROI Detection

Our algorithm uses two regions of interest as first step; i.e., eyes region and face region. These regions are chosen as they conform to large spatial head turns and they have higher probability of being uniquely detected in a video sequence. In literature, there exist many algorithms to detect the face and eyes regions robustly and accurately (see [11] for more details). For current work, we use Haar-like features to detect face and facial feature regions (eyes) in given frame [8]. Haar-like feature algorithm utilizes a boosted cascade of simple feature classifiers to detect object region in an image. Although these detectors demonstrated high processing speeds, but high detection rate can only be achieved for rather strictly near front faces. This introduces the loss-of-tracking problem when the face rotates at a big angle. To overcome this loss-of-tracking problem, we use Tracking-Learning-Detection (TLD) framework [9], a tracking framework that explicitly decomposes the long-term tracking task into tracking, learning and detection. The TLD method is designed for long-term tracking of arbitrary objects in unconstrained environment. The TLD framework improves the tracking performance by combining a detector and an optical-flow tracker. It works well in extreme poses and in occlusion. The problem with TLD is that it needs manual initialization by creating a bounding box using mouse. To this end, Haar algorithm is employed to overcome this shortcoming as an initialization step. Hence, our method estimate the initial location of eyes and face using Haar-like features in a first frame, then the bounding box of these regions are provided to TLD for tracking purposes in successive frames.

2.2 Estimation of Pivot Point

According to facial animation parameters by MPEG-4, pivot point (3D origin of face) for yaw and pitch movement are assumed to be at C2 of the spinal column (cervical vertebra) based on anthropometric knowledge (as shown in fig. 2). Formally, if $O_{yaw} = \{X_{O_{yaw}}, Y_{O_{yaw}}, Z_{O_{yaw}}\}$ are the pivot point of yaw movement at C2 of the spinal column. The values of $X_{O_{yaw}}$, $Y_{O_{yaw}}$ and $Z_{O_{yaw}}$ can be estimated as;

$$\begin{aligned} X_{O_{yaw}} &= FW/2 \\ Y_{O_{yaw}} &= 0 \\ Z_{O_{yaw}} &= 1290ES \end{aligned}$$

Similarly, if $O_{pitch} = \{X_{O_{pitch}}, Y_{O_{pitch}}, Z_{O_{pitch}}\}$ be the pivot

point of pitch movement at C2 of the spinal column. The values of X_{Opitch} , Y_{Opitch} and Z_{Opitch} can be estimated as;

$$\begin{aligned} X_{Opitch} &= FW/2 \\ Y_{Opitch} &= FH/2 \\ Z_{Opitch} &= 1239ES \end{aligned}$$

Where FW and FH are face width and face height respectively and ES equals eye separation distance divides by 1024.

Fig.2. From left to right - For head orientation ‘yaw’ & ‘pitch’ calculated based on the position of human cervical vertebrae depicted in red [10] and according to facial animation parameters by MPEG-4 [6].

2.3 Pose Recovery

For pose recovery, the left-eye center $E_{Left} = \{X_{Left}, Y_{Left}, Z_{Left}\}$, right-eye center $E_{Right} = \{X_{Right}, Y_{Right}, Z_{Right}\}$, eyes midpoint $E_{Mid} = \{X_{Mid}, Y_{Mid}, Z_{Mid}\}$, yaw pivot point $O_{yaw} = \{X_{Oyaw}, Y_{Oyaw}, Z_{Oyaw}\}$ and pitch pivot point $O_{pitch} = \{X_{Opitch}, Y_{Opitch}, Z_{Opitch}\}$ for each frame i are provided by Haar-like features and TLD framework. It is then modeled as the eyes movement with respect to pivot points and calculate yaw, pitch and roll as follows: For yaw angle estimation at frame i , if vector U_i is from yaw pivot point $O_{iyaw} = \{X_{iOyaw}, Y_{iOyaw}, Z_{iOyaw}\}$ to eyes midpoint $E_{iMid} = \{X_{iMid}, Y_{iMid}, Z_{iMid}\}$ in three-dimensional space,

$$U_i = (X_{iMid} - X_{iOyaw}, Y_{iMid} - Y_{iOyaw}, Z_{iMid} - Z_{iOyaw})$$

By projecting U_i (3D vector) to XZ plane (2D vector) constitute another vector Y_i .

$$Y_i = (X_{iMid} - X_{iOyaw}, 0, Z_{iMid} - Z_{iOyaw})$$

The angle α_i between Y_i and Z-axis gives Yaw angle for frame i , as shown in fig. 3 (a). Formally α_i can be derived as;

$$\alpha_i = \arctan(X_{iMid} - X_{iOyaw} / Z_{iMid} - Z_{iOyaw})$$

For pitch angle estimation at frame i , let vector V_i is from pitch pivot point $O_{pitch} = \{X_{Opitch}, Y_{Opitch}, Z_{Opitch}\}$ to eyes midpoint $E_{iMid} = \{X_{iMid}, Y_{iMid}, Z_{iMid}\}$ in three-dimensional space,

$$V_i = (X_{iMid} - X_{iOpitch}, Y_{iMid} - Y_{iOpitch}, Z_{iMid} - Z_{iOpitch})$$

By projecting V_i (3D vector) to YZ plane (2D vector) constitute another vector P_i .

$$P_i = (0, Y_{iMid} - Y_{iOpitch}, Z_{iMid} - Z_{iOpitch})$$

The angle β_i between P_i and Z-axis gives Pitch angle for frame i , as shown in fig. 3(b). Formally β_i can be derived as;

$$\beta_i = \arctan(Y_{iMid} - Y_{iOpitch} / Z_{iMid} - Z_{iOpitch})$$

Similarly for roll angle estimation a 2D vector R_i is constructed from left eye center $E_{Left} = \{X_{Left}, Y_{Left}\}$ to right eye center $E_{Right} = \{X_{Right}, Y_{Right}\}$ in XY-plane is given as,

$$R_i = (X_{iRight} - X_{iLeft}, Y_{iRight} - Y_{iLeft})$$

The angle γ_i between R_i and X-axis gives Roll angle for frame i , as given below;

$$\gamma_i = \arctan(Y_{iRight} - Y_{iLeft} / X_{iRight} - X_{iLeft})$$

The α_i , β_i and γ_i give estimated angles for yaw, pitch and roll respectively.

Fig.3. From left to right- Face geometry employed for Yaw angle and Pitch angle calculation.

3. Experimental Results

To evaluate the effectiveness and accuracy of the

proposed method, we compared pose values (Yaw, Pitch, Roll) of our algorithm with pose values of head mounted inertial measurement unit(IMU) as ground truth. We have selected IMU as a ground truth because almost all available head pose data set on internet are annotated with IMU. In Fig. 4, a person is wearing a helmet containing IMU for ground truth values and face and eyes regions are selected for our approach to get actual pose values. The evaluation is conducted during real-time by simultaneously recording and comparing values of IMU with the pose values from our approach.

The experiment was conducted in which subject appears to be sitting at a distance of about 2 feet from the camera and is asked to move freely within the range of $[-70: +70]$, $[-45: +45]$ and $[-45: +45]$ for yaw, pitch and roll, respectively. The ground truth values and algorithm data are being recorded real-time. The experimental results of yaw, pitch and roll for 120 frames are shown in figure 5(a, b, c). The mean errors and standard deviations for yaw, pitch and roll angles are shown in table 1.

4. Concluding Remarks

In this paper we propose a novel and simple method for head pose estimation using geometric facial regions in a video sequence. Proposed geometric method employs anthropometric knowledge of face to create model for head motion as rigid body movement by introducing pivot points for yaw and pitch. Our algorithm employs TLD tracker for tracking facial region even under occluded environment. We divided the computation of yaw, pitch and roll into separate estimation of orientation about Y, X and Z-axis, respectively. It can be seen from the results that our method is comparable and, in case performs identical to head motion estimation based on IMU. Our algorithm can simultaneously estimate yaw, pitch and roll angles in each frame with accuracy.

Fig.4. Experimental Setup used for ground truth calculation based on IMU mounted on the subject's head.

Fig. 5. (a,b,c): From Top to Bottom- Estimated pose angles and corresponding ground truth for Yaw, Pitch and Roll.

Furthermore, the advantage of the proposed scheme is that no prior knowledge regarding camera parameters and distance of the user from the camera is needed. This work has successfully solved the problem of head pose using monocular camera under un-intrusive and uncontrolled environment.

Table1. Mean errors & stand. dev. for three rotation angles.

Rotation Angles	Mean Error	Std. dev.
Yaw	1.3	1.16
Pitch	5.225	4.62
Roll	4.41	3.43

References

- (1) Li, H., Roivainen, P., Forchheimer, R. : “3-D motion estimation in model-based facial image coding”, *IEEE Pattern Analysis and Machine Intelligence* 15, 545-555, 1993.
- (2) Lablack, A.; Maquet, F., Ihaddadene, N. & Djeraba, C.: “Visual gaze projection in front of a target scene”, *IEEE International Conference on Multimedia and Expo (ICME)*, New York City, NY - USA, 2009.
- (3) Valenti, R., Sebe, N., Gevers, T.: “Combining Head Pose and Eye Location Information for Gaze Estimation”, *IEEE Tran. on Image Processing* 21, 802-815, 2012.
- (4) Rudovic, O., Pantic, M., Patras, I.: “Coupled Gaussian Processes for Pose-Invariant Facial Expression Recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (in press), 2012.
- (5) Murphy-Chutorian, E., Trivedi, M.: “Head Pose Estimation in Computer Vision: A Survey”, *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 607—626, 2009.
- (6) Tekalp, A., Ostermann, J.: “Face and 2-D mesh animation in MPEG-4”, *Signal Processing: Image Communication* 15, 387-421, 2000.
- (7) Marieb, E.: “Essential of human anatomy and physiology”, San Fransisco, USA: Pearson, 150 –151, 2006.
- (8) Chen, D., Huaqiao, U., Liu, QZ.: “Generalized Haar-Like Features for Fast Face Detection”, *Machine Learning and Cybernetics* 4, 2131-2135, 2000.
- (9) Kalal, Z., Mikolajczyk, K., Matas, J.: “Tracking – Learning- Detection”, *Pattern Analysis and Machine Intelligence*, *IEEE Trans.* 34, 1409-1422, 2012.
- (10) Hawkes, B.: “Cervical Fusion Mid Cervical. Morphopedics. Where Technology and Orthopedics Collides”. accessed jan 2012. (2012).
- (11) Zhang, C., Zhang, Z., “A Survey of Recent Advances in Face Detection”. Microsoft Research-Technical Report-MSR-TR-2010-66. (2010).
- (12) Adel Lablack, Jean Martinet and Chabane Djeraba: “Head Pose Estimation Using a Texture Model Based on Gabor Wavelets”, *Pattern Recognition Recent Advances*, Adam Herout (Ed.), ISBN: 978-953-7619-90-9, InTech, 2010.
- (13) Jeffrey Ng and Shaogang Gong: “Multi-View Face Detection and Pose Estimation Using A Composite Support Vector Machine across the View Sphere”, In *Proc. IEEE International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, 14-21, 1999.
- (14) Huang, J., Shao, X. & Wechsler, H.: “Face pose discrimination using support vector machines (SVM)”, *14th International Conference on Pattern Recognition (ICPR) Brisbane - Australia*, Vol 1, 154-156, 1998.
- (15) L. Chen, L. Zhang, Y. Hu, M. Li, H. Zhang: “Head Pose Estimation using Fisher Manifold Learning”, In *IEEE International Workshop on Analysis and Modeling of Faces and Gesture*, 203-207, 2003.
- (16) Srinivasan, S. & Boyer, K.: “Head-pose estimation using view based eigenspaces”, *16th International Conference on Pattern Recognition (ICPR)*, Quebec City – Canada, 2002.
- (17) Vineeth N. Balasubramanian, Sreekar Krishna, Sethuraman Panchanathan: “Person - independent head pose estimation using biased manifold embedding”, In *EURASIP J. Adv. Signal Process*, Vol. 2008 (January 2008).
- (18) Ruigang Yang: “Model-based head pose tracking with stereovision”, In *Proceedings of Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, 255—260, 2002.
- (19) Stephen Milborrow & Fred Nicolls: “Locating Facial Features with an Extended Active Shape Model”, In *Proceedings of the 10th European Conference on Computer Vision: Part IV*, 504 – 513, 2008.
- (20) T. F. Cootes, G. J. Edwards & C.J. Taylor : “Active Appearance Models”. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 484—498, 1998.
- (21) D. Cristinacce and T. Cootes: “Feature detection and tracking with constrained local models”, In *BMVC 2006*, pages 929—938, 2006.
- (22) T. Vatahska, Maren Bennewitz, Sven Behnke: “Feature -based head pose estimation from images”, *Humanoids*, 330-335, 2007.
- (23) A. H. Gee and R. Cipolla: “Determine the gaze of faces in images”, *Image and Vision Computing*, 12(10), 639–647, 1994.
- (24) R. Stiefelhagen, J. Yang, and A. Waibel: “A model-based gaze tracking system”, In *Proc of the IEEE Int. Joint Symp. on Intelli. and Sys.*, 1996.
- (25) T. Horprasert, Y. Yacoob, and L. Davis, “Computing 3-D Head Orientation from a Monocular Image Sequence,” *Proc. IEEE Int’l Conf. Automatic Face and Gesture Recognition*, pp. 242-247, 1996.