IDIAP COMMUNICATION

A Video Database for Head Pose Tracking Evaluation

Silèye O. Ba [a] Jean-Marc Odobez [a]

IDIAP–Com 05-04

September 2005

[a] IDIAP Research Institute

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny − Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

1 Introduction

The automatic analysis of human behavior constitutes a rich research field. It is required for computers in smart houses, car or meeting rooms to interact smoothly with people. Automatic human behavior analysis can be based on speech with speech recognition or keyword spotting based systems. Images or video sequences can be used to analyze non verbal behavior such as hand pointing, visual focus of attention, body and head gestures. Head related behaviors such as head gestures and visual focus of attention are an important subgroup of the non verbal behaviors. Thus tracking the head and estimating its pose is crucial in many computer vision applications. Many investigations have been made in the field of head tracking and pose estimation. Without being exhaustive we can cite [1, 2, 3, 4, 5, 6, 7, 8, 9]. As good automatic analysis of head related behaviors will be based on good head tracking and pose estimation, the performances of the proposed methods have to be rigorously evaluated. To evaluate the performances of head pose tracking algorithms head pose databases are required. Many still head pose images database exist such as the FERET database [10], the PIE database [11] and the Prima-Pointing 04 database [12]. Still head pose image databases are very useful to build head pose models and evaluate head pose detection or recognition algorithms but tracking algorithms require head poses in real life video sequences. As head pose annotation in video sequences is not a trivial task, most of the time, head pose tracking algorithms are evaluated qualitatively on video sequences without head pose annotations. Some people used head pose video databases to evaluate their algorithms [9] but their database is not publicly available. In all the cases, comparing head pose tracking performances is difficult. As people are working on head pose tracking since a long time, a publicly available to evaluate tracking and pose estimation algorithms and compare performances is required. This work makes some steps in this direction.

This document describes our work to provide a video database, of people in real situation with their head pose continuously annotated through time. To build a head pose database, first a head pose representation has to be selected then a procedure to annotate head pose defined. In our case, head poses were annotated using a magnetic 3d location and orientation tracker, the flock of bird [13]. The environment of our recordings were a meeting room and an office with their common light sources. These environments summarize well indoor environments for head pose tracking. The recording in the meeting room involved 16 persons, 2 persons per meeting lasting each approximatively 10 minutes. The office recording involved also 15 persons, 1 person per recording lasting each approximatively 8 minutes. The procedure to acquire a copy of the database can be obtained by sending an email to databases@idiap.ch.

The remaining of this document is organized as followed, Section 2 describes two ways to represent head pose the first one base one the camera frame representation and the second one based on the head frame representation. Section 3 describes the way our database was built and the information it contains. Section 5 gives the conclusions.

2 Head Pose Representation

A head pose representation can be defined as a parameterization of the head rotations with respect to a given frame. Thus, many representation are possible. To each reference frame correspond a possible head pose representation. A common representation is the Euler angles parameterization. Among the multiple possible Euler angle parameterization two are commonly used. The first one use as reference frame the camera frame. This representation was used to build the PIE database. A second parameterization use as reference frame a frame attached to the head. This representation was used to build the Prima-Pointing database. In this Section we describe these two head pose representations and gives ways to pass from a representation to another.

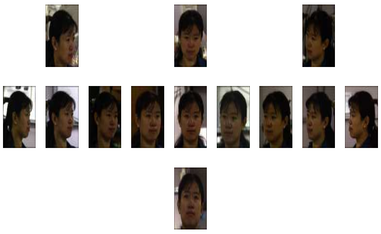

Figure 1: Head poses in the PIE representation (from PIE database)

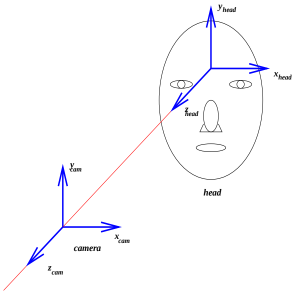

Figure 2: Frontal head pose configuration

2.1 The PIE Representation

The setup to build the PIE database was the following, 9 of the 13 cameras was positioned at a roughly head height in an arc from approximately a full left profile to a full right profile. Each neighboring pair of these 9 cameras are therefore approximately 22.5 degrees apart. Of the remaining 4 camera 2 were placed above and below the central (frontal) camera and 2 were placed in the corners of the room. Image of the views of the 13 camera are given in Figure 1.

in the PIE representation, the pose of a person is defined relatively to a reference configuration between the head and the camera. When the reference configuration head-camera is realized the person is said to be in a frontal pose. If we defined a basis $\mathcal{B}_{head}$ rigidly attached to the head such that the x axis is parallel to the line of the eyes, the y axis is parallel to the line of the nose and the z axis is orthogonal x and y axis, the frontal pose corresponds to the head-camera configuration in which the head reference $\mathcal{B}_{head}$ is up to a translation, aligned to the camera reference $\mathcal{B}_{cam}$ (see Figure 2).

In PIE Representation, a head pose defining a given head-camera configuration is determined by three Euler angles (α, β, γ). These angles defines three consecutive rotations that, when applied, align the basis $\mathcal{B}_{head}$ in the frontal head-camera configuration (frontal pose) to the basis $\mathcal{B}_{head}$ in the given head-camera configuration. The first rotation is about an angle β around the axis y_c of the camera, the second one is about an angle α around x_c of the camera, and the third one about an angle γ around the axis z_c of the camera. This sequence of rotations defines a rotation matrix $M_{\alpha,\beta,\gamma}$. α is called head tilt, β head pan γ is called head roll. $M_{\alpha,\beta,\gamma}$ can be written, using the rotation matrixes $M_{x_c,\alpha}$, $M_{y_c,\beta}$ and $M_{z_c,\gamma}$ of the rotation matrixes about angles α, β and γ around the axes x_c , y_c and z_c:

$$M_{\alpha,\beta,\gamma} = M_{z_c,\gamma} M_{x_c,\alpha} M_{y_c,\beta} \tag{1}$$

More details about rotations representations and Euler angles parameterization can be found in [14].

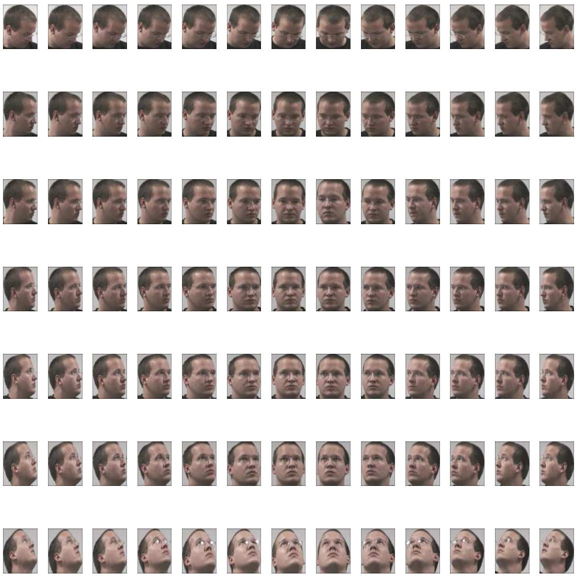

Figure 3: Head poses in the POINTING representation (from Prima-Pointing database)

2.2 The POINTING Representation

This representation has been used to build the Prima-Pointing database. Figure 3 shows the full pose range of a person in this database. In this representation, a basis is also rigidly attached to the head and, as for the PIE representation, the frontal head pose is defined by the head-camera configuration in which the head and the camera bases (B_{head} and B_{cam}) are aligned.

A head pose is defined by three Euler angles (α, β, γ) representing three consecutive rotations that align $\mathcal{B}_{head}$ of the frontal configuration to $\mathcal{B}_{head}$ in the head pose configuration. But, in the POINTING representation, the rotations are done with respect to the axes of $\mathcal{B}_{head}$ the basis rigidly attached to the head instead of $\mathcal{B}_{cam}$. The angles (α, β, γ) correspond to the classical Euler angles. For our database, the three consecutive rotations are the following. First a rotation about γ around z_h axis, then a rotation about α around y_h axis and finally a rotation about β around the axis x_h. The rotation matrixes $M_{\alpha,\beta,\gamma}$ describing these three consecutive rotations can be written, using the rotation matrixes $M_{y_h,\alpha}$, $M_{xh,\beta}$ and $M_{z_h,\gamma}$ of the rotation matrixes about angles α, β and γ around the axes y_h , x_h and z_h:

$$M_{\alpha,\beta,\gamma} = M_{z_h,\gamma} M_{y_h,\beta} M_{x_h,\alpha}$$

2.3 From a Representation to Another

Each one of the two representations have advantages and drawbacks. In a perceptive point of view, the POINTING representation is more natural. Perceptually, the PIE Representation can is seen as if the head was static and the camera rotating around it. But, in this representation, the head roll for given pan and tilt corresponds an image in plane rotation. More precisely given the image appearance of the head pose $(\alpha, \beta, 0)$, the image appearance of the head pose (α, β, γ) corresponds just to an in plane rotation of angle γ of the image appearance corresponding to $(\alpha, \beta, 0)$. This property does not not hold for the POINTING representation. Thus, because depending on the cases people may be interested in using one of the representations, being able to convert the head pose in a given representation into the head pose of the other representation is useful.

If we denote (α, β, γ) and $M^*_{\alpha,\beta,\gamma}$ the head pose and its corresponding rotation matrix in the PIE representation and $(\alpha', \beta', \gamma')$ and $M^+_{\alpha',\beta',\gamma'}$ the same head pose in the single POINTING representation. Thus given (α, β, γ) (resp $(\alpha', \beta', \gamma')$) it's corresponding head pose in the POINTING (resp. PIE) representation is obtained by solving the equation $M^{+}_{\alpha',\beta',\gamma'} = M^{*}_{\alpha,\beta,\gamma}$ to find $(\alpha', \beta', \gamma')$ (resp. (α, β, γ)).

3 Set-up and Recordings

3.1 Head Pose Annotation with The Flock of Bird

We used the Pointing Representation to build our video head pose database. We used two cameras, in for the office recording and another one in the meeting room recording. The office camera was recording at 12.5 frames per second and the meeting room camera at 25 frames per second. The cameras was fixed, and head poses are defined with respect to a reference head-camera configuration. The cameras were calibrated using the methodology described in [15]. A matlab camera calibration toolbox using this methodology is downloadable from [16]. The outputs of the calibration are the intrinsic and the extrinsic parameters of the camera. The external camera parameters define an affine a translation $t_{ext,cam}$ and rotation $R_{ext,cam}$ relating any point X^{cam} the camera basis $\mathcal{B}_{cam}$ and its representation X^{ext} in the external basis $\mathcal{B}_{ext}$ by the formula $X^{cam} = R_{ext,cam}X^{ext} + t_{ext,cam}$.

For the head pose annotation we used a device the flock of bird (FOB) [13]. The FOB is a 3D location and orientation tracker with mainly two components. A reference basis rigidly attached to the desk $\mathcal{B}_{fob}$ and, a bird rigidly attached to the head of a person defining the head basis $\mathcal{B}_{head}$. The FOB outputs locations and orientations with respect to its reference basis $\mathcal{B}_{fob}$. Thus, we needed to find a transformation to give the FOB outputs with respect to the camera basis. We called this procedure calibration of the FOB to the camera. If we denote $\{P_i,\ i = 1, ..., N\}$ the coordinates of a set of points into the external basis $\mathcal{B}_{ext}$. Lets also denote by $\{P_i^{fob},\ i = 1, ..., N\}$ the coordinates of the same points in the FOB basis $\mathcal{B}_{fob}$. The FOB-camera calibration corresponds to find a translation $t_{fob,ext}$ and rotation $R_{fob,ext}$ such that the coordinates of a point in the external frame X^{ext} are obtained from the coordinates of the point in the FOB frame X^{fob} by the relation $X^{ext} = R_{fob,ext}X^{fob} + t_{fob,ext}$. The rotation matrix $R_{fob,ext}$ and the translation vector $t_{fob,ext}$ can be approximated numerically by solving the optimization problem:

$$(R_{fob,ext}, t_{fob,ext}) = \arg\min_{R,t} \sum_{i=1}^{N} ||RP_i^{fob} + t - P_i||_2^2 \tag{2}$$

Thus, the affine transform to pass from the FOB basis to the camera basis is defined by the translation vector $t_{fob,cam} = R_{ext,cam}t_{fob,ext} + t_{ext,cam}$ and the rotation matrix $R_{fob,cam} = R_{ext,cam}R_{fob,ext}$.

If $(\alpha^0_{fob}, \beta^0_{fob}, \gamma^0_{fob})$ is the frontal head pose and with respect to the FOB basis $\mathcal{B}_{fob}$. The head pose corresponding to a FOB output $(\alpha_{fob}, \beta_{fob}, \gamma_{fob})$ can be obtained using $R_{fob,cam}$ by finding (α, β, γ) such that:

$$M_{\alpha,\beta,\gamma} = R_{fob,cam} M_{\alpha_{fob},\beta_{fob},\gamma_{fob}} M^T_{\alpha^0_{fob},\beta^0_{fob},\gamma^0_{fob}} R^T_{fob,cam} \tag{3}$$

where M^T denotes the transposition of the matrix M, which correspond to its inverse for rotation (orthogonal) matrixes.

In our recording setup the camera and the FOB were set on manually. There was a time delay between the recording starting time of the devices. The FOB outputs and the video frames have to be aligned. For the alignment it is necessary to find at list one easy-to-identify event in the video sequence and in the FOB data. We defined the alignment event to be quick head shake. This gesture corresponds to an oscillation of the head pan while the head tilt and roll are steady in the FOB data. The video frames corresponding to the peaks of this oscillation are easy to find in the video sequence also. The time instants $\{t_i^{ref,fob}\ i = 1, ..., N_t\}$ of the peaks in the FOB data and the time instants $\{t_i^{ref,cam}\ i = 1, ..., N_t\}$ of the corresponding video frames in the camera recording will be used as

Office	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
length (minutes)	4.6	4	5.5	4.3	5.6	7	6.2	5.5	5.6	5.6	7.2	7.1	5.4	6.3	6.7

Table 1: Office recording lengths in minutes

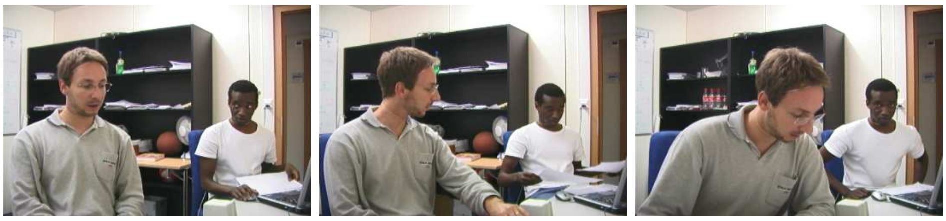

Figure 4: Images from the first office recording (Office 1)

reference time to compute the delay D between the starting time of the two devices:

$$D = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(t_i^{ref,cam} - t_i^{ref,fob} \right) \tag{4}$$

In theory D could be estimated with only the time instant corresponding to one peak, using the time instants of many peaks reduce the noise effect. The relation between a camera instant t^{cam} and a FOB instant t^{fob} is given by $t^{cam} = t^{fob} + D$. From this relation we can derive the correspondence between the frames of the video and the FOB.

3.2 Office Recordings

The Office recordings involved 15 persons, the length of each recording is given in Table 1. In each recording, a person was sitting in front of a computer and acting in the framework of a simple scenario with the three following parts:

- look at the camera in a frontal head pose and perform an alignment gesture
- look at fixed points of the room
- interact with the experimenter

The office recording set up was close to human computer interaction set up. The head image sizes were quite high resolution varying approximately between 100 × 100 and 180 × 180.

For each video sequence, the flock recordings, after alignment and transformations using the FOB-camera calibration procedure, give the head pose annotations. Figure 5 shows the distribution of the head pan, tilt, and roll values for the whole office recordings. Pan value are ranging from -50 to 150 in a quasi uniform distribution. The tilt values are ranging from -60 to 20 while most of the values are within -15 and 15. Roll values are ranging from -20 to 20 while the values are concentrated between -5 and 10. Figure 6 displays the scatter plot of the head pan versus head tilt in the first office recording (Office 1) recording. In this plot, we can notice that the pan are mostly positive. The reason is that the experimenter (person in the background in Figure 4) was sitting on the side of the positive pan (right side) of the annotated person.

3.3 Meeting Room recordings

In the meeting room, 8 meetings were recorded. In each meeting, 4 persons were involved and among them, two had their head pose continuously annotated. The durations of the meeting are given in table 2. These meeting were recorded according to a simple scenario. The people had to

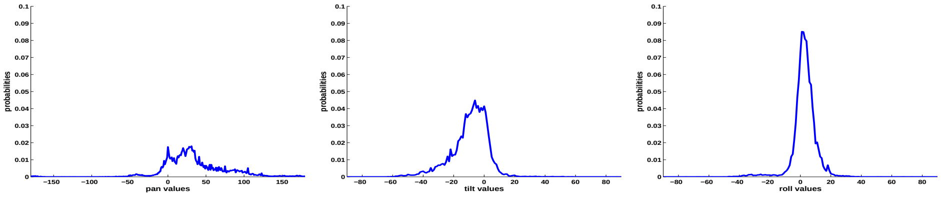

Figure 5: Distribution of the pan (left), tilt (center), and roll (right) angles in the office recordings

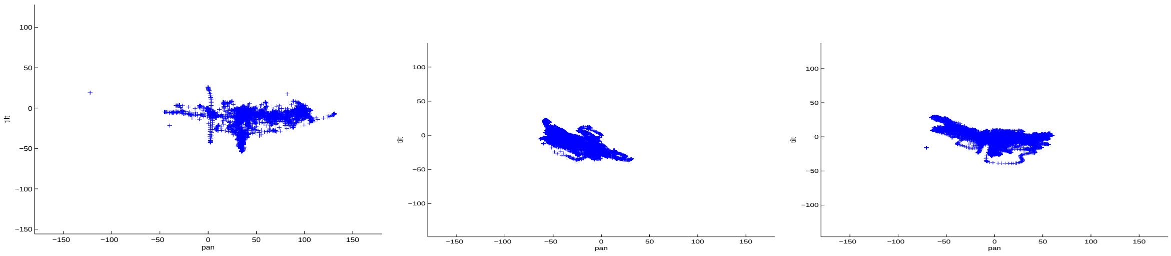

Figure 6: Typical pan versus tilt plot. Left: office recording (Office 1). Center: meeting 1 person Right. Right: meeting 1 Person Left

- look at the camera in a frontal head pose (in some cases perform a FOB video alignment gesture)
- write their name on a sheet of paper
- discuss statements displayed on the projection screen

This scenario gives a lot of freedom to the participants and people were acting very naturally as in real situations. The size of head images in the recordings was approximately 100 × 100.

Figure 8 gives the distribution of head pan, tilt, and roll angles over the meeting head pose dataset. The pan values are ranging within -90 and 50 with two modes, -50 and 10 degrees. The tilt values are ranging from -50 to 20 and the roll values from -20 to 40 degrees. Figure 6 display the pan versus tilt scatter plots of the two persons having their head pose annotated. It can be seen in this figure that people's pan values were often negative. Negative pan values corresponds mainly to looking at the projection screen which was an important visual focus of attention for the persons according to our scenario.

4 Head Pose Tracking Evaluation

To evaluate a head pose tracking system two kind of errors have to be evaluated: head pose estimation errors and head localization errors.

4.1 Head Pose estimation Evaluation

A head pose defines a vector in the 3D space, the vector indicating where the head is pointing at. It can be thought of as a vector based on he center of the head and passing through the nose. This vector depends only on the head pan and tilt values in the Pointing representation. The angle between the 3D pointing vectors defined by the head pose ground truth and the pose estimated by the tracker can be used as a first pose estimation error measure. This error measure is well suited for studies of visual focus of attention where the main concern is to know where the head/person is looking at. However it

Meeting	1	2	3	4	5	6	7	8
length (minutes)	7.6	7.5	7.3	11	10	10.9	14.3	12.1

Table 2: Meeting room recording lengths in minutes

Figure 7: Images from a Meeting 1: person right and person left

gives no information about the roll estimation errors. In order to have more details about the origin of the errors we will measure the individual errors made separately on the pan, tilt and roll in the Pointing representation.

For each one of the four errors we will compute the mean, the standard deviation and the and the median value of the absolute errors. We use the median value because it is les sensitive to extremal values than the mean. Thus the median will be less biased than the mean by short term period pose estimation errors due to bad head localization. These errors measures were used with a portion of the meeting recordings to evaluate and compare head pose tracking performances of four algorithms in [2]

4.2 Head Localization Evaluation

At each time t, if we denote $GS(t)$ the the area of the box locating the head of a person and $TS(t)$ the area of the head box estimated by a a tracker. The tracking localization errors can be measured by $e(t) = \frac{1}{2}\left(\frac{|GS(t) \cap TS(t)|}{|GS(t)|} + \frac{|GS(t) \cap TS(t)|}{|TS(t)|}\right)$. This error is 0 when tracking is perfect, and 1 when it totally fails. This error measure for head localization is simple. More sophisticated localization errors can be found in [17]. It is worth noticing that in our database, the head location ground truth is not available. Only is given the FOB position which is located near one of the ears of the corresponding person.

5 Conclusion

In this paper we describe the building steps of the IDIAP Head Pose Database. The database contains video sequences of people with their head pose continuously annotated. The recording were done in a meeting room and an office in very natural conditions. There were no restriction in people's motions or poses. To make some steps towards common evaluation database for algorithms in the head tracking and pose estimation community we made the database is publicly available. This database can also be used to study problem such as pose based head gesture modeling, facial expression analysis. Information for it's acquisition can be obtained by sending an email to databases@idiap.ch.

Acknowledgment

The authors want to thanks **Rutger Rienks** who kindly provides the 3D magnetic field location and orientation tracker from his Lab (Electro-Technic Department, University of Twente) in a collaborative

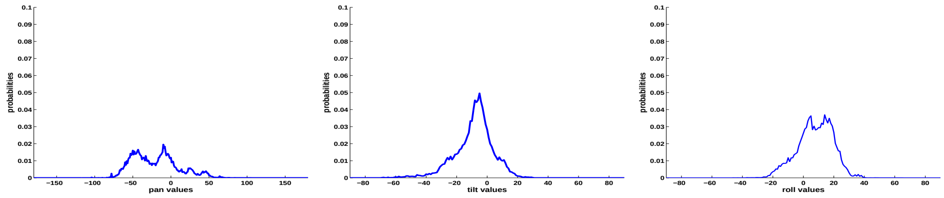

Figure 8: Distribution of head pan (left), tilt (center), and roll (right) angles in the meeting recordings

work between IDIAP and University of Twente within the Integrated Project AMI (Augmented Multi-party Interaction).

References

[1] L. Brown and Y. Tian, “A study of coarse head pose estimation,” in *Proc. of IEEE Workshop on Motion and Video Computing*, Orlando Florida, Dec 2002, pp. 183–191.

[2] S.O. Ba. and J.M. Odobez, “Evaluation of head pose tracking algorithms in natural environments,” in *Proc. of International Conference of Multimedia & Expo (ICME)*, Amsterdam, Netherland, Aug 2005, pp. 264–267.

[3] T. Cootes and P. Kittipanya-ngam, “Comparing variations on the active appearance model algorithm,” in *Proc. of British Machine Vision Conference (BMVC)*, Norwich UK, Sept 2002.

[4] B. Kruger, S. Bruns, and G. Sommer, “Efficient head pose estimation with gabor wavelet,” in *Proc. of British Machine Vision Conference (BMVC)*, Bristol UK, Sept 2000.

[5] R. Rae and H. Ritter, “Recognition of human head orientation based on artificial neural networks,” *IEEE Trans. on Neural Network*, vol. 9(2), pp. 257–265, March 1998.

[6] R. Stiefelhagen, J. Yang, and A. Waibel, “Estimating focus of attention based on gaze and sound,” in *Proc of Workshop on Perceptive User Interface (PUI)*, Florida, USA, Nov 2001, pp. 1–9.

[7] S. Niyogi and W. Freeman, “Example-based head tracking,” in *Proc. of Int. Conf. on Automatic Face and Gesture Recognition (ICAFGR)*, Killington, Vermont USA, Oct 1996, pp. 374–377.

[8] R. Yang and Z. Zhang, “Model-based head pose tracking with stereo vision,” Tech. Rep., Microsoft Research, Oct 2001.

[9] Y. Wu and K. Toyama, “Wide range illumination insensitive head orientation estimation,” in *Proc. of IEEE Conf. on Automatic Face and Gesture Recognition (AFGR)*, Grenoble France, Apr 2001, pp. 183–188.

[10] P.J. Phillips, P.J. Rauss H. Moon, and S. Rizvi, “The feret evaluation methodology for face recognition algorithms,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 22(10), pp. 1090–1104, Oct 2000.

[11] T. Sim and S. Baker, “The cmu pose, illumination, and expression database,” *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 25(12), pp. 1615–1618, Oct 2003.

[12] “Pointing’04 icpr workshop: Head pose image database,” .

[13] “Flock of birds,” http://www.ascension-tech.com/products/flockofbirds.php.

[14] M.K. Spong and M. Vidyasagar, *Robot dynamics and control*, John Wiley & Sons, USA, first edition, 1998.

[15] Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientation," in *Proc. of International Conference on Computer Vision (ICCV)*, San Diego, CA, Sept 1999.

[16] "Camera calibration toolbox for matlab," http://www.vision.caltech.edu/bouguetj/calib_doc/.

[17] K. Smith, D. Gatica-Perez, J. Odobez, and S. Ba, "Evaluating multi-object tracking," in *Proc. of IEEE Workshop on Empirical Evaluation Methods in Computer Vision (EEMCV)*, June 2005.