Head Pose Estimation on Low Resolution Images

Nicolas Gourier, Jérôme Maisonnasse, Daniela Hall, James L. Crowley

PRIMA, GRAVIR-IMAG
INRIA Rhône-Alpes,
38349 St. Ismier.
France.

Abstract. This paper addresses the problem of estimating head pose over a wide range of angles from low-resolution images. Faces are detected using chrominance-based features. Grey-level normalized face imagettes serve as input for linear auto-associative memory. One memory is computed for each pose using a Widrow-Hoff learning rule. Head pose is classified with a winner-takes-all process. We compare results from our method with abilities of human subjects to estimate head pose from the same data set. Our method achieves similar results in estimating orientation in tilt (head nodding) angle, and higher precision for estimating orientation in the pan (side-to-side) angle.

1. Introduction

Knowing the head pose of a person provides important cues concerning visual focus of attention [12]. Applications such as video surveillance, intelligent environments and human interaction modelling require head pose estimation from low-resolution face images. Unfortunately, most methods described in the research literature require high-resolution images, often using multiple views of the face. In this paper we address the problem of estimating head pose from low-resolution single images.

The pose, or orientation, of a head is determined by 3 angles: slant, pan and tilt. The slant angle represents the person's head inclination with regard to the image plane, whereas the tilt and the pan angles represent the vertical and the horizontal inclination of the face. Our objective is to obtain a reliable estimation of head pose on unconstrained low-resolution images.

We employ a fast, chrominance-based segmentation algorithm to isolate and normalize the face region in size and slant. We then project this region of the image into a small fixed-size imagette using a tranformation that normalises size and slant orientation. Normalised face imagettes are used to train an auto-associative memory using the Widrow-Hoff correction rule. Classification of head pose is obtained by comparing normalised face imagettes with those reconstructed by the auto-associative memory. The head pose which obtains the highest score is selected. We compare results with our method to human performance on head pose estimation using the same data set [13]. This process is described in section 3. We compare results from this method with human performance for head pose estimation using the same data set, as described in section 4. Results of these comparisons are discussed in section 5.

2. Approaches to Head Pose Estimation

Local or global approaches exist for head pose estimation. Local approaches usually estimate head pose from a set of facial features such as eyes, eyebrows and lips. Three dimensional rotation of the head can be estimated from correspondences between such facial landmarks in the image and the face [1], [2], [3]. However, the detection of facial features tends to be sensitive to partial changes of illumination, person and pose variations. Robust techniques have been proposed to handle such variations [4], [5] but these require high resolution images of the face and tracking can fail when certain facial features are occluded. Some local-based systems, such as FaceLAB [21], have a precision smaller than one degree. Such systems use stereo vision and require high resolution of the image of the face. Transformation-based approaches use some geometric properties of facial landmarks to estimate the 3D rotation of the head [6], [7], [8]. However, such techniques remain sensitive to the precision of detected regions and to the resolution of the face image. Such problems do not appear when using global approaches.

Global approaches use the entire image of the face to estimate head pose. The principal advantage of global approaches is that only the face needs to be located. No facial landmark, or face model are required. Global approaches can accommodate very low resolution images of the face. Template matching is a popular method to estimate head pose. The best template is found via a nearest-neighbour algorithm, and the pose associated with this template is selected as the best pose. Template matching can be performed using Gabor Wavelets and Principle Components Analysis (PCA) [9], or Support Vector Machines [10], but these approaches tend to be sensitive to alignment and are dependent on the identity of the person. Neural networks have also been used for head pose estimation [11]. Stiefelhagen [12] reports 10 degrees of precision on the Pointing'04 Head Pose Image Database [13]. However, some images of the same users were used both in training and testing. Furthermore, the number of cells in hidden layers is chosen arbitrarily, which prevent creation of image class prototypes.

In the method described in this paper, we adapt auto-associative memories based on the Widrow-Hoff learning rule. Auto-associative memories require very few parameters and contain no hidden layers [14]. Prototypes of image classes can be saved and reused. The Widrow-Hoff learning rule provides robustness to partial occlusions [22]. Each head pose serves to train an auto-associative network. Head pose is estimated by selecting the auto-associative network with the highest likelihood score.

3. Head Pose Estimation using Linear Auto-associative Neural Networks

3.1 Linear Auto-associative Memories

Linear auto-associative memories are a particular case of one-layer linear neural networks where input patterns are associated with each other. Auto-associative memories associate images with their respective class, even when the image has been degraded or partially occluded. With this approach, each cell corresponds to an input pattern. We describe a grey-level input image x' with a normalized vector x = x'/||x'||. A set of M images composed of N pixels of the same class are stored into a N x M matrix X = ($x_{1,}$ x_2, …, x_M). The linear auto-associative memory is represented by a connection matrix W.

The reconstructed image y_k is obtained by computing the product between the source image x and the connection weighted matrix W_k : $y_k = W_k \cdot x$. The similarity between the source image and a class k of images is estimated as the cosine between x and y_k: cos(x, y) = $x \cdot y^T$. A similarity of 1 corresponds to a perfect match. The connection matrix W_k^0 is initialized with the standard Hebbian learning rule $W_k^0 = X_k \cdot X_k^T$. Reconstructed images with Hebbian learning are equal to the first eigenface of the image class. To improve recognition abilities of the neural network, we learn W with the Widrow-Hoff rule.

3.2 The Widrow-Hoff Correction Rule

The Widrow-Hoff correction rule is a local supervised learning rule. At each presentation of an image, each cell modifies its weights from the others. Images X, of the same class are presented iteratively with an adaptation step η until all images are classified correctly. As a result, the connection matrix W_k becomes spherically normalized. The Widrow-Hoff learning rule is described by:

$$W_k^{t+1} = W_k^t + \eta \cdot (x - W_k^t \cdot x) \cdot x^T \qquad (1)$$

In-class images are minimally deformed by multiplying with the connection matrix, while extra-class images are more strongly deformed. Direct comparison between input and output normalized images gives a score between 0 and 1. This correction rule has shown good results on classic face analysis problems in the case of images from a single camera, for problems such as face recognition, sex classification and facial type classification [14].

The Widrow-Hoff correction rule increases the performance of PCA and provides robustness to partial occlusions [22]. All dimensions are used and few parameters are needed. There is no requirement to specify the choice of the structure or the number of cells in hidden layers. Furthermore, prototypes, W_k, of image classes can be saved, recovered and directly reused on other images unlike non-linear memories or neural networks with hidden layers, where prototypes can not be recovered.

3.3 Head Pose Image Database

The choice of a good database is crucial for learning. The Pointing'04 Head Pose Image database [13] consists of 15 sets of images of different people. Each set contains 2 series of 93 images of the same person at different poses. Subjects are 20 to 40 years old. Five people have facial hair and seven are wearing glasses.

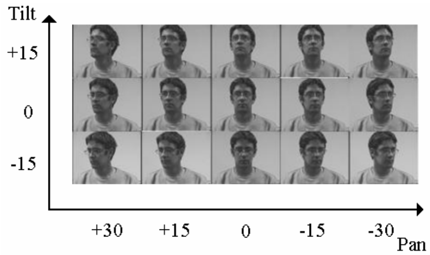

Figure 1. Sample of the Pointing'04 Image Database

Head pose is determined by pan and tilt angle. Each angle varies between -90 and +90 degrees, with a step of 15 degrees for pan, and 30 and 15 for tilt. Negative values for tilt correspond to bottom poses and positive values correspond to top poses. During the database acquisition, people were asked to look successively at 93 markers. Each marker corresponds to a particular pose. A sample of the database can be seen Figure 1.

3.3 Head Pose Prototypes

The face region is normalized into a low resolution normalized grey-scale imagette of 23x30 pixels, as in [4]. Face normalization provides invariance of position, scale and slant [15]. This increases the reliability of results and processing time becomes independent of original face size. All further operations take place within this imagette.

We consider each head pose as a class. A connection matrix W_k is computed for each pose k. The Pointing'04 database consists in 13 Poses for pan and 9 Poses for tilt. Two experiments have been performed using this approach. In the "separate" technique, we learn each angle on an axis while varying the angle of the other axis. Each classifier corresponding to a pan angle is trained with varying tilt angle. Similarly, each memory corresponding to a tilt angle is trained with a varying pan angle. The "separate" experiment learns 22 classifiers: Pan = +90,…, Pan = -90, Tilt = +90,…, Tilt = -90. We use an adaptation step η of 0.008 for pan and 0.006 for tilt for this experiment. Pan and tilt are trained separately.

In the "grouped" experiment, pan and tilt angle are trained together. Each classifier corresponds to a pan and a tilt angle. This experiment learns 93 classifiers: (Pan,Tilt)=(0,-90),…, (Pan, Tilt)=(+90,+75), (Pan, Tilt)=(0,+90). We use an adaptation step η of 0.006 for this experiment.

To estimate head pose on a given face imagette, a simple winner-takes-all process is employed. We compute the cosine between the source image X and reconstructed images X_k'. The pose whose memory obtains the best match is selected (2).

$$Pose = \arg\!-\max_k(\cos(X, X'_k)) \qquad (2)$$

4. Human Abilities for Head Pose Estimation

To our knowledge, there is no data avalaible concerning human abilities for estimating head pose from images. Kersten [18] reports that front and profile poses are particularly well recognized by humans. These poses are used as key poses [19]. This observation is true not only for head poses, but also for other objects. However, they do not estimate intermediate poses.

As a comparison to our artificial system, we measured the performance of a group of 72 human subjects on head pose estimation. In our experiment, we have tested 36 men and 36 women, ranging in age from 15 to 60 years old. The experiment consisted of two parts: one for pan angle estimation, and the other for tilt angle. Images from the Pointing'04 Head Pose Database were presented in random order to the subject for 7 seconds, with a different order for each subject. Subjects were asked to examine the image, and to select an answer pose estimation from a fixed set. The data base consists of 65 images for pan and 45 for tilt, which gives 5 images for each pose.

The psycho-physical basis for human head pose estimation from static images is unknown. We do not know whether humans have a natural ability to estimate head pose from such images, or whether people must be trained for this task using annotated images. In order to avoid bias in our experiment, the subjects were divided into 2 groups: people in the first group may inspect the labelled training images of head pose as long as they wish before beginning the experiment, whereas people in the second group are not provided an opportunity to see the images before the experiment. First and second groups are respectively referred as "Calibrated" and "Non-Calibrated" subjects. Creating these two groups allows us to determine if training significantly increases human performances on head pose estimation.

5. Results and Discussion

In this section, we compare results of the two variations of our method (separate and grouped) using the Pointing'04 Head Pose image database. There are two ways of splitting the data for training and testing. By using the first set of the database as the training data and testing on the second set, we measure the performance of our system on known users. By using the Jack-Knife method, also known as the leave-one-out algorithm, we measure the performance on unknown users. To have an idea of the efficiency of our system in human-computer interaction applications, we compare performances of our system with human performances.

5.1 Evaluation Measures

To evaluate the performance of our system, we must define evaluation criteria. Average absolute error for pan and tilt is the main evaluation metric. It is computed by averaging the difference between expected pose and estimated pose for all images. We also compute average absolute error for pan and tilt per pose. The Pointing'04 database is well suited for such measure, because it provides the same amount of data for each pose. Precise classification rate and correct classification with 15 degrees errors is also computed. We compare results of our system on known and unknown users. Results are presented in Table 1.

5.2 Performances

Our system works well with known subjects on both angles. With the separate technique, we achieve a mean error of 7.3 degrees in pan and 12.1 degrees in tilt. The grouped learning provides a mean error of 8.5 degrees in pan and 10.1 degrees in tilt. Pan angle can be correctly estimated with a precision of 15 degrees in more than 90% of cases with both learning techniques.

Results obtained with the Jack-Knife algorithm show that our system also generalizes well to previous unseen subjects and is robust to identity. With the separate technique, we see that pan angle is well recognized with an average error of 10.3 degrees. Average error decreases to 10.1 degrees using the grouped learning. The average tilt error is 15.9 degrees using the separate technique, and 16.8 degrees using the grouped technique. Average error per pose is shown in Figure 2.

Known Users	LAAM separate	LAAM grouped
Pan Average Error	7.3 º	8.5 º
Tilt Average Error	12.1 º	10.1 º
Pan Class. With 0º	61.3 %	60.8 %
Tilt Class. With 0º	53.8 %	61.7 %
Pan Class. With 15º	93.3 %	90.1 %

Unknown Users	LAAM separate	LAAM grouped
Pan Average Error	10.3 º	10.1 º
Tilt Average Error	15.9 º	16.8 º
Pan Class. With 0º	50.4 %	50 %
Tilt Class. With 0º	43.9 %	44.5 %
Pan Class. With 15º	88.1 %	88.7 %

Table 1. Performance evaluation on known and unknown users. LAAM refers to linear auto-associative memories

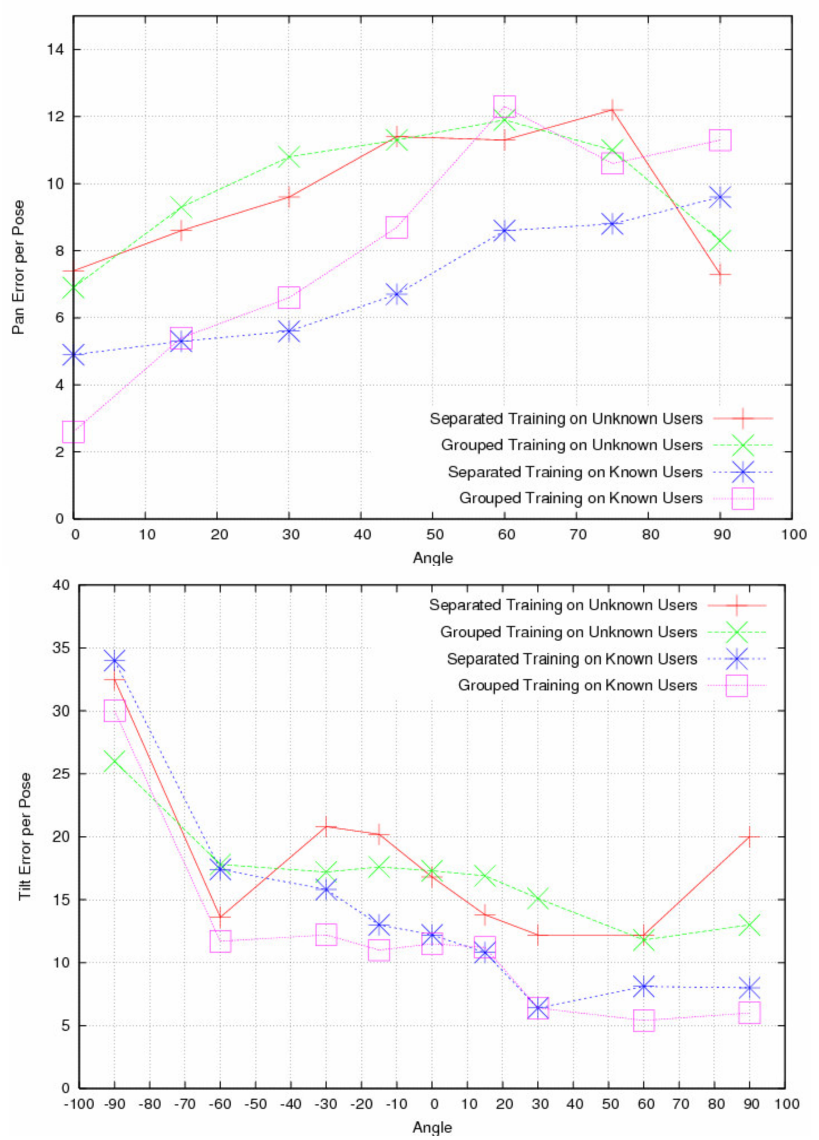

Figure 2. Average error per pose for known and unknown users

Concerning the pan angle, the average absolute error in pose is relatively stable with both techniques. Grouped and separated training accommodate intermediate tilt angles. We achieve a 0 degrees classification rate of 50.4% for pan angle and 21 % for tilt angles with the separate technique. Using the grouped technique provides a

50% classification rate for pan angle and 45% for tilt angle. Pan angle can be correctly estimated with a precision of 15 degrees in 88% of cases. These results tend to show that using the together technique does not provide significantly improve results. Examples can be seen in Figure 4.

Faces are not aligned in the Pointing'04 database. Normalizing face images provides small variations in alignment. Results show that our system can handle alignment problems. Computing a score for each memory allows us to discriminate face and non-face images. Head detection and pose estimation is done in a single process. The system runs at 15 images/secs using the separate technique, and 3 images/secs with the grouped technique. As humans estimated angles separately, we will use the separate learning for comparison with human performances.

5.3 Comparison to Human Performances

We computed the same evaluation measures for humans. Results for calibrated (C) and non-calibrated (NC) people are shown in Table 2. Global human average error for head pose estimation is 11.9 degrees in pan and 11 degrees in tilt. Creating two groups allows comparing the performances of our system on unknown users to the best human performances. We apply a Student test to compare the two populations. Calibrated people do not perform significantly much better in pan. However, the difference is significant in tilt angle. These results show that pan angle estimation appears to be natural for humans, whereas tilt angle estimation is not. This is due to the fact that people twist their head left and right more often than up and down during social interactions. In situations when people talk to each other, pan angle provides good cues on visual focus of attention [12], [19]. Head poses changes in tilt become meaningless. This is even more relevant when people are seated, because their head is roughly at the same height. People are more used to consider pose changes in pan. Seeing training images annotated do not improve much pan angle estimation but improves significantly tilt angle estimation. The best human performance is obtained by calibrated people.

	C	NC	LAAM Sep. U
Pan Average Error	11.8 °	11.9 °	10.3 °
Tilt Average Error	9.4 °	12.6 °	15.9 °
Pan Class. with 0 °	40.7 %	42.4 %	50.4 %
Tilt Class. With 0°	59 %	48 %	43.9 %

Table 2. Human/Machine performance evaluation. C and NC stand for Calibrated and Non-Calibrated people

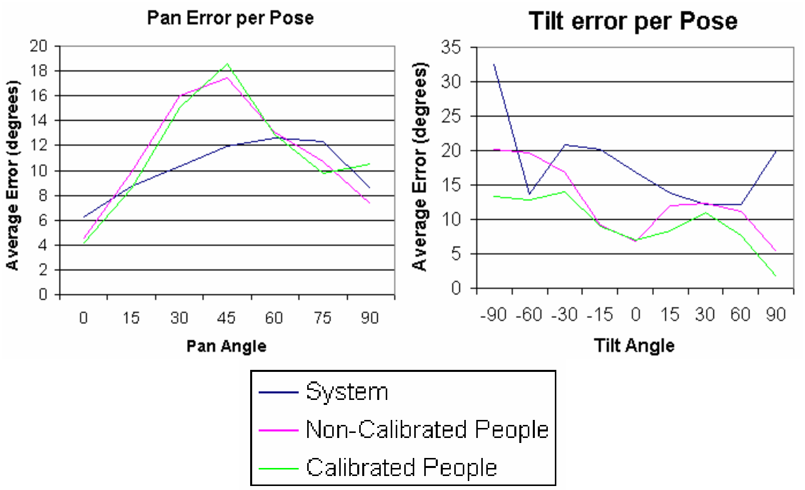

Figure 3. Human / System performance per pose

Average error per pose for human subjects can be seen in Figure 3. For pan angle, we found that humans perform well for front and profile angles, but not for intermediate angles. The average error per pose in pan can be modelled by a Gaussian function centered at 45 degrees. Minimum error can be found at 0 degrees, which corresponds to front pose. Furthermore, during our experiment, we observe that most people did not use intermediate angles such as 30, 45 and 60 degrees. These results suggest that the human brain uses front and profile as key poses, as suggested in [17]. Concerning tilt angle, humans performs better for top angles than for bottom angles. The minimum error can be found at +90 degrees, whereas the maximum error is at -90 degrees. This can be due to the fact that when a face is nodding downward, hair dominates a large surface of the apparent face, providing more information about side to side angle.

With an average error of 10.3 degrees and a precise classification rate of 50.4%, our method performs significantly better than humans at estimating pan angle (11.9 degrees). The standard deviation of the average error per pose is low for the system and high for humans. The system achieves roughly the same precision for front and profile, and higher precision for intermediate poses. With an average error of 11 degrees, humans perform better in tilt angle. Our method performs well for top poses. This is due to the fact that hair becomes more visible in the image and the face appearance between people changes more when looking down. On the other hand, such changes are less visible for up poses.

Face region normalization also introduces a problem. The height of the neck changes between people. This provides high variations on face imagettes and can disrupt tilt angle estimation.

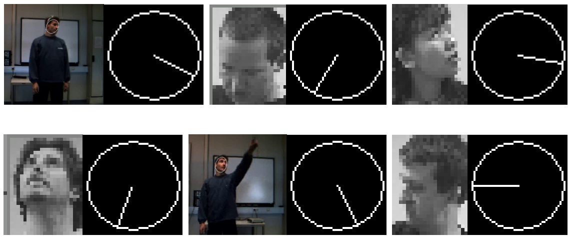

Figure 4. Pan angle estimation on example images

6. Conclusion

We have proposed a new method to estimate head pose on unconstrained low resolution images. Face image is normalized in scale and slant into an imagette by a robust face detector. Face imagettes containing the same head pose are learned with the Widrow-Hoff correction rule to obtain a linear auto-associative memory. To estimate head pose, we compare source and reconstructed images using their cosine. A simple winner-takes-all process is applied to select the head pose which prototype gives the best match.

We achieved a precision of 10.3 degrees in pan and 15.9 degrees in tilt only on unknown subjects on the Pointing'04 Head Pose Image database. Learning pan and tilt together does not provide significantly better results. Our method provides good results on very low resolution face images and can handle wide movements, which is particularly adapted to wide-angle or panoramic cameras setups. The system generalizes well to unknown users, is robust to alignment and runs at 15 frames/secs.

We measured human performance on head pose estimation using the same data set. Our system performs significantly better than humans in pan, especially with intermediate angles. Humans perform better in tilt. Results of our system may be improved by fitting an ellipse to delimit more precisely the face. Our head pose estimation system can be adapted to video sequences for situations such as human interaction modelling, video surveillance and intelligent environments. By knowing a coarse estimate of the current head pose, the temporal context can help to limit head pose search only to neighbour poses. The use of head prototypes reduces significantly the computational time in video sequences.

7. References

[1] A.H. Gee, R. Cipolla, "Non-intrusive gaze tracking for human-computer interaction," Mechatronics and Machine Vision in Practise, pp. 112-117, 1994.

[2] R. Stiefelhagen, J. Yang, A. Waibel, "Tracking Eyes and Monitoring Eye Gaze," Workshop on Perceptual User Interfaces, pp. 98-100, Banff, Canada, October 1997.

[3] A. Azarbayejani, T. Starner, B. Horowitz, A. Pentland, "Visually Controlled Graphics," IEEE Transactions on PAMI 15(6) 1993, pp. 602-605.

[4] N. Gourier, D. Hall, J. Crowley, "Estimating Face Orientation using Robust Detection of Salient Facial Features," Pointing 2004, ICPR, Visual Observation of Deictic Gestures, Cambridge, UK.

[5] J. Wu, J.M. Pedersen, D. Putthividhya, D. Norgaard, M.M. Trivedi, "A Two-Level Pose Estimation Framework Using Majority Voting of Gabor Wavelets and Bunch Graph Analysis," Pointing 2004, ICPR, Visual Observation of Deictic Gestures, Cambridge, UK.

[6] Q. Chen, H. Wu, T. Fukumoto, M. Yachida, "3D Head Pose Estimation without Feature Tracking," AFGR, April 16/1998 Nara, Japan. pp. 88-93.

[7] R. Brunelli, "Estimation of Pose and Illuminant Direction for Face Processing," Proceedings of IVC(15), No. 10, October 1997, pp. 741-748.

[8] P. Yao, G. Evans, A. Calway, "Using Affine Correspondance to Estimate 3-D Facial Pose," 8th ICIP 2001, Thessaloniki, Greece, pp. 919-922.

[9] S. McKenna, S. Gong, "Real-time face pose estimation," *International Journal on Real Time Imaging*, Special Issue on Real-time Visual Monitoring and Inspection, volume 4: pp.333-347, 1998.

[10] J. Ng, S. Gong, "Multi-view Face Detection and Pose Estimation using a Composite Support Vector Machine across the View Sphere," International Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, p. 14-21, Corfu, Greece, September 1999.

[11] B. Schiele, A. Waibel, "Gaze tracking based on face-color," Workshop on Automatic Face and Gesture Recognition, pages 344-349, Zurich, June 26-28, 1995.

[12] R. Stiefelhagen, "Estimating Head Pose with Neural Networks - Results on the Pointing04 ICPR Workshop Evaluation Data," Pointing 2004, ICPR, Visual Observation of Deictic Gestures, Cambridge, UK.

[13] N. Gourier, J. Letessier, "The Pointing 04 Data Sets," Pointing 2004, ICPR, Visual Observation of Deictic Gestures, Cambridge, UK.

[14] D. Valentin, H. Abdi, A. O'Toole, "Categorization and identification of human face images by neural networks: A review of linear auto-associator and principal component approaches," *Journal of Biological Systems* 2, pp. 413-429, 1994.

[15] K. Schwerdt, J. Crowley, "Robust face Tracking using Color," International Conference on Automatic face and Gesture Recognition pp. 90-95, 2000.

[16] G.J. Klinker, S.A. Shafer, T. Kanade, "A Physical Approach to Color Image Understanding," International Journal on Computer Vision 1990.

[17] H. Abdi, D. Valentin, "Modeles Neuronaux, Connectionistes et Numeriques de la Reconnaissance des Visages," Psychologie Francaise, 39(4), pp. 357-392, 1994.

[18] D. Kersten, N.F. Troje, H.H. Bülthoff, "Phenomenal competition for poses of the human head," Perception, 25 (1996), pp. 367-368.

[19] B. Steinzor, “The spatial factor in face to face discussions,” *Journal of Abnormal and Social Psychology* 1950 (45), pp. 552-555.

[20] H.H. Bülthoff, S.Y. Edelmann, M.J. Tarr, “How are three-dimensional objects represented in the brain?,” *Cerebral Cortex 1995* (5) 3, pp. 247-260.

[21] Seeing Machines Company. “FaceLAB4”, *http://www.seeingmachines.com*

[22] H. Abdi, D. Valentin, "Modeles Neuronaux, Connectionistes et Numeriques de la Reconnaissance des Visages," Psychologie Francaise, 39(4), pp. 357-392, 1994.