A New Head Pose Estimating Algorithm based on a Novel Feature Space for Driver Assistant Systems

Ali Ghaffari[#1], Mahdieh Rezvan[#2], Alireza Khodayari*[3], Seyyed Hossein Sadati*[4], Afra Vahidi-Shams[#5]
[#] Islamic Azad University, Tehran South Branch, Tehran, Iran
[1]ghaffari@kntu.ac.ir, [2]mhrezvan@gmail.com, [5]afra.vahidishams@gmail.com
[*] K. N. Toosi University of Technology, Tehran, Iran
[3]arkhodayari@ieee.org, [4]sadati@kntu.ac.ir,

***Abstrac*—This paper introduces a Direct Accountability technique for real-time estimation of head's position and orientation. Unlike existing works which rely on feature extraction either in the image domain or in 3D space, our proposed approach based on expert classifiers ordering estimates the head pose for real-time applications, such as human-machine interaction (HMI) and video-based surveillance systems. In this paper a method for feature extraction is proposed which aims to keep only the angle related features that are independent of identity. Moreover a performance study is provided aiming to evaluate the accuracy of the proposed approach. Experimental results show that the proposed features could describe the pose of driver face image successfully, and have a good performance in real scenes specially for automobile safety and security and driver assistant devices.**

Keywords—*Head pose estimation, Human-Machine Interface, intelligence Automotive Systems, Driver Behaviours Surveillance*

I. INTRODUCTION

Human face detection and recognition techniques play important roles in applications like video surveillance, human computer interface and recently human-robot interaction. Head poses are important indicators of a person's focus of attention. Determining a face pose is a complex problem, and numerous methods have been proposed for its estimation [1, 2]. A historical view of the driver assistant systems shows a rapid development of these systems. In recent years, several techniques for vision pose estimation have been proposed [3-6]. The main application domain was advanced driver assistance.

In the machine vision, head pose estimation is the process of orientation of a human head. An ideal head pose estimator, like other facial vision processing steps, must demonstrate invariance to various factors as well as biological appearance. These factors include physical phenomena like camera distortion, lighting, facial expression, and the presence of accessories like glasses and hats. The proposed approaches can be broadly classified into two main categories: 3D [7] and 2D head pose estimation approaches [8]. The facial features detection and tracking in video sequences are two of the main challenging problems in machine vision. This research area has many applications such as gaze detection, teleconferencing, etc. [9]. In general after detecting the face in a frame, automatic detection of facial features in video sequences is performed. Many approaches have been proposed to detect faces in static scenes, such as Principal Component Analysis [10, 11], clustering and neural nets [12]. These techniques obtained good detection accuracy rate, but they are computationally expensive and hardly achieve real-time performance.

In designing our method, the system should be able to estimate head pose from a single camera. Also the system should be able to estimate a continuous range of head orientation with fast (30fps or faster) operation. And finally, the system must work regardless of the specific driver and operating conditions. We are interested in developing intelligent interfaces and this system is the first step to give computers the ability to look at people, which may be a good way to improve human-computer interaction.

In recent years, there have been various approaches in literature for head pose estimation. Existing methods can be described in following main categories that suggest approaches which have been used for head positions estimations [13]:

-*Appearance Template Methods* compare a new image of a head to a set of exemplars (each labeled with a discrete pose) in order to find the most similar view.

-*Detector Array Methods* train a series of head detectors each attuned to a specific pose and assign a discrete pose to the detector with the greatest support.

-*Nonlinear Regression Methods* use nonlinear regression tools to develop a functional mapping from the image or feature data to a head pose measurement

-*Geometric Methods* use the location of features such as the eyes, mouth, and nose tip to determine pose from their relative configuration.

-*Tracking Methods* recover the global pose change of the head from the observed movement between video frames.

-*Hybrid Methods* combine one or more of these aforementioned methods to overcome the inherent limitations in any single approach.

The remainder of this paper is organized as follows. Introduction reviews some techniques related to our work. The next section presents the methodology of the work. In this Section, the developed algorithm for face detection and tracking is presented. The proposed methodology and

results are presented in Sections 4, respectively. The last section presents the conclusions and scopes for future studies.

II. DETAILED DESIGN AND IMPLEMENTATION

Our aim is estimating the head pose parameters from the video sequences. In other words, we track the stereo head pose over time. In this section, we propose an approach that creates a feature space from the video sequences which only contains angle related features. Figure 1 shows the algorithm flow.

Our framework consists of these stages: 1. Image enhancement, 2. Face detection, 3. Features Extraction, 4. Pose estimation using classification. These steps are illustrated in the following sections.

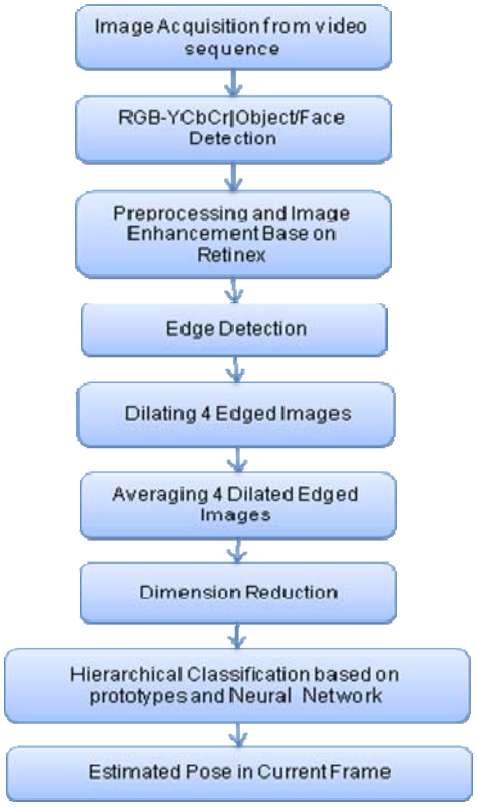

Figure 1. The column The flow of the proposed algorithm

A. Improving Algorithm

The light condition is the main restriction in our work and must be normal. In other word, the faces which are going to be detected are too bright or too dark. We increased the robustness of light variations by developing a light compensation / correction preprocessing technique. We used "Retinex" which is a well-known algorithm of image enhancement. The retinex is aimed to obtain the balance between the human vision and machine vision system along with color constancy [14]. The first idea of Retinex was proposed by Land [15] as a model of lightness and color perception of the human vision. Obviously it is not only a model, but also could be developed to algorithms of image enhancement. After Land, Jobson and his coworkers [16] defined a single-scale Retinex (SSR), which is an implementation of center/surround Retinex. The Single-scale retinex is given by:

$$R_i(x,y) = \log I_i(x,y) - \log[F(x,y) * I_i(x,y)] \quad (1)$$

Where Ii(x,y) is image distribution in the ith color band, F(x,y) is the normalized surround function.

$$\iint F(x,y)dxdy = 1 \quad (2)$$

The image distribution is the product of scenes reflectance and illumination.

$$I_i(x,y) = S_i(x,y)r_i(x,y) \quad (3)$$

Where $S_i(x,y)$ is the spatial distribution of illumination and $r_i(x,y)$, the distribution of scene reflectance. Various surround functions could be used. We used the Gaussian formula with absolute parameter c,

$$F(x,y) = \exp(-r^2 / c^2) \quad (4)$$

But depending on the special scale, SSR can either provide dynamic range compression (small scale) or tonal rendition (large scale). Superposition of weighted different scale SSR is a choice to balance these two effects which are provided by multi scale Retinex (MSR):

$$R_{MSRi} = \sum_{n=1}^{N} \omega_n R_{ni} \quad (5)$$

Where N is the number of the scales, R_{ni} is the i_{th} component of the n^{th} scale. The obvious question about MSR is the number of scales needed, scale values, and weight values. Experiments showed that three scales are enough for most of the images, and the weights can be equal. Generally fixed scales of 15, 80 and 250 can be used, or scales of fixed portion of image size can be used. But these are more experimental than theoretical, because we do not know the scale of image in the real scenes. The weights can be adjusted to weight more on dynamic range compression or color rendition [17].

B. Face Detection

We used edge detection and then filled holes to separate objects in each frame and estimate face-like regions. At this stage, sometimes there are several face-like regions in the image block. We selected the area with maximum pixels as the face region. With 1200 test images of 15 subjects, background complexities and lighting conditions, the correct face detection rate is found to be 86% in this work. It is worth mentioning that the proposed system does not recognize whether the identified face region is accurate enough for next stages.

C. Extracting Features

At previous stages, a face detector is applied to detect the facial region. In this stage edge detection in four directions is performed on the images, and then the images after edge dilation are averaged together. The result image is the pose feature space which is used for classification. The feature spaces are shown in figure 2 for a number of subjects with different poses.

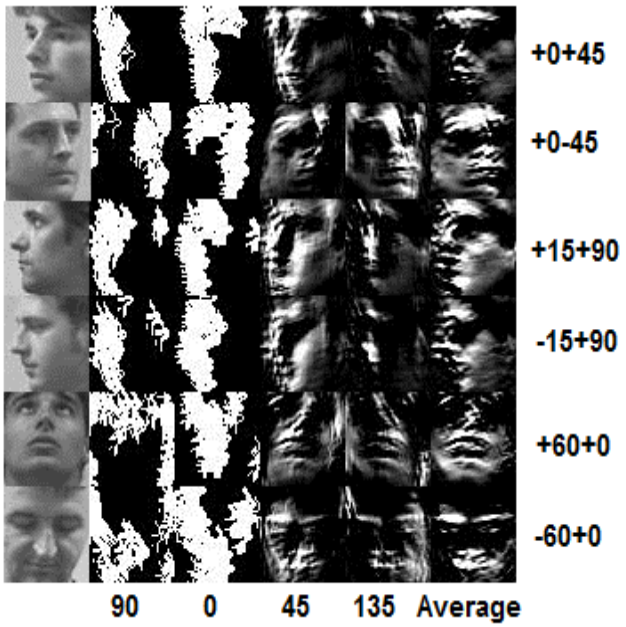

Figure 2. The feature spaces for some different subjects in various angles

D. Classification

The next stages in our method are prototype fabrication, the definition of a calibration method, and then study of a tracking framework for the driver's gaze. We normalized feature vector from previous step and calculated the average of training images to create the prototype for each face pose, also these features are stored in the new dataset so that the features of new frame can be compared by one nearest neighbor classifier.

III. Experimental Setup and Results

We implement our feature space for head pose estimation and present the results on the public head pose database (Pointing'04 database [18]).

A. Dataset

The Pointing'04 database consists of 15 sets of images, wearing and not wearing glasses and having various skin colors. Each set contains 2 series of 93 images of the same person, and the 93 head poses are determined by yaw and tilt degrees, which vary from -90° to +90°. One subject with various head poses is shown in figure 3.

Figure 3. Sample images from the Pointing'04 head pose data base

B. Cropping based on nose-tip Method

We first cropped each image based on nose-tip method. To reduce the influence of the background, the image patches are not cropped centering on the nose-tips. Instead, the location of the nose-tip in the cropped image is shifted from the center according to the known head poses so that the background can be excluded as much as possible [19]. As shown in figure 4, by specifying the nose-tip location (x_n,y_n) and the ground truth pan and tilt angle (θ,φ), the center of the image patch(of size W × H) determined by:

$$x_c = x_n + p\frac{W_\Phi}{180} \tag{6}$$

$$y_c = y_n + p\frac{H_\Phi}{180} \tag{7}$$

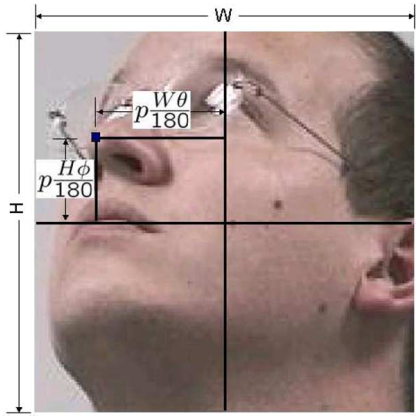

Figure 4. Cropping image patch according to the location of the nose-tip and head pose

C. Edge Detection

In this step the face cropped images should be changed so that to be independent of identity features and only the features which are related to the angle remain. Such a picture that encompasses only the face features related to the angle constitutes our feature space. The edge detection process is performed by convolving the image with an edge operator (a 2-D filter known as mask). In this work, Horizontal and vertical Sobel operators and also Robert's operator in 45 and 135 directions have been used.

D. Edge dilation on edge detected images

Considering that the edges can change depending on various environmental conditions such as lighting and partial head displacement in different images of the same angle, which can make different results in edge detection, a method is needed to dilate the edges in the edge detected image. For this purpose a random window of size 6×6 is selected from the parts of image which include edge information and convolved with the images resulted from each edge detectors. This process is repeated several times by choosing different windows from different parts of the image. Then obtained images are averaged together. With this action edges will be dilated on the image. Finally, four images resulted from any of operators after dilation are put together and averaged pixel by pixel.

E. Classification

The aim of this section is Pattern Recognition, which means recognizing and classifying the object based on its

specificities. At the end of this stage the process of pose estimating will be finished. In this article MLP neural network is used for classification.

Considering the large number of classes in the case of head pose estimation on Pointing04 database, in the following experiments the problem space are divided into smaller spaces to increase the accuracy of the estimation.

Experiment One) The first experiment is performed on the images that just contained pan changes with 13 angles and the tilt angles remained unchanged. The purpose of this experiment is the review of the separation ability of the extracted features by the presented method in this article just with considering the pan classification. The test included seven stages that each of them corresponded to one of the tilt angles. Each stage contained 13 angles. So the problem is a thirteen-class problem.

For solving this thirteen-class problem first, according to what was explained in method part, after making feature space from the test image by edge detection, dilation, and dimension reduction done by PCA, the classification process is performed. For training the classifier, the images of eight subjects are used as the training set and the images of seven subjects as the test set, which all have been selected randomly. Figure 8 shows the result of this experiment achieved from 20 times run. The experiment was repeated by 10 and 12 training samples. The results have been shown in figure 5.

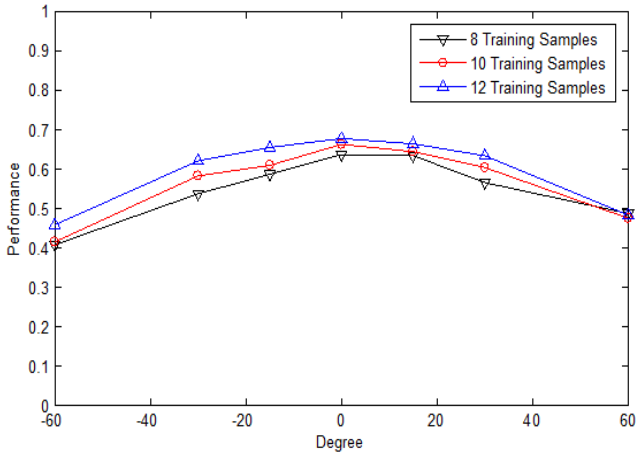

Figure 5. Classification results in the horizontal direction with a number of training samples

Experiment Two) In the next test, images are used that just changed in tilt angles and did not include pan changes, to evaluate the efficiency of our feature extraction technique with just tilt classification in fixed pan angles. The experiment is done in 13 stages that each of them corresponded to one of the pan angles and contained seven angles of tilt with this fixed pan angle. So, in this case the problem is a seven-class problem.

To resolve this problem first feature space is made from the test image by edge detection, dilation and averaging, like the steps described in the first experiment. Then the dimension is reduced by means of PCA and finally is sent for classification. The number of training and test images selected for this experiment, respectively are eight and seven which are randomly selected. The Picture below shows the results of this test performed in 20 times run. Results were repeated for 10 and 12 training image samples that are respectively shown in Figure 8 in 20 times run.

Experiment Three) In this experiment for head pose estimation the method is applied on whole database space with a different classification. In this test, images are classified by two separate classifiers in pan and tilt direction. Each of these two MLP networks conducted the classification process parallel to another on the feature space created from images after reducing the dimension by PCA. The first MLP in tilt direction is a seven-class classifier and the next one in pan direction is a thirteen-class classifier. The end outcome is the presented angle for the test image. Result of this experiment for 20 times run is about 36.73% by using the images of eight subjects as training set and the images of seven subjects as test set, which are shown in table1.

The experiment was also repeated by 10 and 12 training samples and the results are illustrated in figure 6 separately.

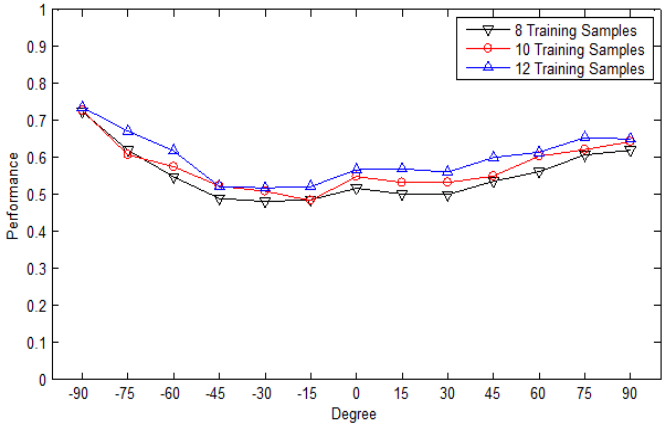

Figure 6. Classification results in the vertical direction with a number of training samples

F. Experiments for Real-Time Video

We have also performed an experiment to evaluate the performance of our method for estimating the head poses from sequences in video format of the data base. Table I. shows the performance evaluation and comparison of the model with some models presented before. The table illustrates that on the whole the results of our proposed method are better than the methods based on PNMF [11], NMF [11], PCA [11], HOSVD [20], LAAM [21] and NN-based [22]. We cannot quantitatively evaluate the results, but most of head poses are estimated very fast and with high accuracy. The experiment shows that our method performance is acceptable and can be used in the real time and hard conditions.

TABLE I.
PERFORMANCE EVALUATION AND COMPARISON TABLE OF THE MODEL

	Experiment3			Comparing Methods					
Metric	Mean	min	max	PNMF [11]	HOSVD [20]	NMF [11]	LAAM [21]	NN-based [22]	PCA [11]
Mean pan err.(Degree)	9.97°	7.7°	12.73°	11.2°	12.9°	10.3°	8.5°	9.5°	13.73°
Mean tilt err.(Degree)	13.78°	11.2°	16.64°	12.8°	17.9°	15.9°	10.1°	9.7°	14.78°
pan Classification	64.67%	61.26%	69.51%		49.25%	50.4%	60.8%	66.3%	55.20%
tilt Classification	57.23%	53.57%	62.91%		54.84%	43.9%	61.7%	52%	57.99%

The time needed by the algorithm is directly related to the image quality and scale of head. The minimum time needed in our work is about 20ms; and maximum about 34ms. The average time needed is 28ms, which means the algorithm has high real time performance.

IV. CONCLUSION

In this work, an effective algorithm for real time head pose estimation based on a novel feature space was proposed. This method is useful in many applications such as driver assistant systems (DAS) and video based surveillance systems. In these applications real-time speed is a critical requirement which our model performs fast with high accuracy. The result can still represent the scale of the driver's head rotation. The focus of this article was whether the driver's attention is disturbed or not. That is when the head direction has been deviated for a long time, so the proposed algorithm provides a judgment based on this task.

The advantages of the proposed technique are as follow. First, the technique of feature extraction in the image domain is independent of identity and only image features related to the angle remain. Furthermore, the technique can handle significant occlusions in which the feature extraction method does not need the appearing of facial components on the sequences as clearly as possible. Thus, our algorithm can detect those incomplete faces resulted from great occlusions and large orientations. The proposed system can be used in Driver Assistant Devices, Collision Prevention Systems and other ITS applications.

REFERENCES

[1] J. Wu, M. Trivedi, "A two-stage head pose estimation framework and evaluation," Pattern Recognition, vol. 41, no. 3, pp. 1138−1158, 2008.

[2] K. Guo, G. Yu, and Z. Li, "A new algorithm for analyzing driver's attention state," Intelligent Vehicles Symposium, 2009 IEEE, pp. 21− 23, June 2009.

[3] M. N. Mamatha, S. Ramachandran ,"Automatic Eyewinks Interpretation System Using Face Orientation Recognition For Human-Machine Interface Orientation Recognition For Human-Machine Interface", International Journal of Computer Science and Network Security (IJCSNS), VOL.9, No.5, May. 2009.

[4] E. Murphy-Chutorian, M. Trivedi, "Hybrid head orientation and position estimation (HyHOPE): A system and evaluation for driver support," in Proc. IEEE Intelligent Vehicles Symposium, 2008.

[5] L. D. Baskar, B. De Schutter, H. Hellendoorn, "Model-Based Predictive Traffic Control for Intelligent Vehicles: Dynamic Speed Limits and Dynamic Lane Allocation", IEEE Intelligent Vehicles Symposium, Eindhoven University of Technology, pp.174-79, 2008.

[6] E. M. Chutorian, M. M. Trivedi, "Head Pose Estimation and Augmented Reality Tracking: An Integrated System and Evaluation for Monitoring Driver Awareness," IEEE Transactions on Intelligent Transportation System, 2010.

[7] J. G. Wang, E. Sung, "EM enhancement of 3D head pose estimated by point at infinity," Image and Vision Computing, vol. 25, no. 12, pp. 1864− 1874, 2007.

[8] J. Xiao, S. Baker, I. Matthews, and T. Kanade, "Real-time combined 2D+3D active appearance models," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, vol. 2, 2004, pp. 535− 542.

[9] A. Doshi, M. M.Trivedi, "On the Roles of Eye Gaze and Head Dynamics in Predicting Driver's Intent to Change Lanes", IEEE Transactions On Intelligent Transportation Systems, Vol. 10, 2009, pp. 453-462.

[10] Y. Li, S. Gong, J. Sherrah, H. Liddell, "Support vector machine based multi-view face detection and recognition," Image and Vision Computing, vol. 22, no. 5, p. 2004, 2004.

[11] X. Liu ,H. Lu ,H. Luo ,"A new representation method of head images for Head Pose Estimation , " IEEE International Conference on Image Processing (ICIP), ,2009.

[12] M. Voit, K. Nickel, R. Stiefelhagen, "Neural network-based head pose estimation and multiview fusion," in Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds., vol. 4122, 2007, pp. 291–298.

[13] E. M. Chutorian, M. M. Trivedi, "Head Pose Estimation in Computer Vision: A Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008.

[14] K. R. Joshi, R. S. Kamathe, "Quantification of Retinex in Enhancement of whether Degraded Image," International Conference on Audio, Language and Image Processing(ICALIP), pp. 1229-1233, 2008.

[15] E. Land, "An alternative technique for the computation of the designator in the retinex theory of color vision", Proc. Nat. Acad, Sci., vol.83, P3078-3080, 1986.

[16] D. J. Jobson, Z. Rahman, and G. A. Woodell, "Properties and erformance of a Center/Surround Retinex," IEEE Transactions on Image Processing, March 1997.

[17] Z. Bian, Y. Zhang, "Retinex image enhancement techniques:algorithm, application and advantages," Final project report for EE264 image processing and reconstruction, 2002.

[18] N. Gourier, D. Hall, and J. Crowley, "Estimating face orientation from robust detection of salient facial structures," in Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures, pp. 17− 25, 2004.

[19] J. Tu, Y. Fu, T. S. Huang, "Locating Nose-Tips and Estimating Head Poses in Images by Tensorposes", IEEE Transactions on circuits and systems for video technology, VOL. 19, NO. 1, pp. 90-102, January 2009.

[20] J. Tu, Y. Fu, Y. Hu, and T. Huang, "Evaluation of head pose estimation for studio data," in Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006, ser. Lecture Notes in

Computer Science, R. Stiefelhagen and J. Garofolo, Eds., vol. 4122, pp. 281–290, 2007.

[21] N. Gourier, J. Maisonnasse, D. Hall, and J. Crowley, "Head pose estimation on low resolution images," in Multimodal Technologies for Perception of Humans, Int'l. Workshop Classification of Events Activities and Relationships, CLEAR 2006, ser. Lecture Notes in Computer Science, R. Stiefelhagen and J. Garofolo, Eds., vol. 4122, pp. 270− 280, 2007.

[22] R. Stiefelhagen, "Estimating head pose with neural networks – results on the Pointing04 ICPR workshop evaluation data," in Proc. Pointing 2004 Workshop: Visual Observation of Deictic Gestures, 2004.