A Natural Head Pose and Eye Gaze Dataset

Stylianos Asteriadis
stiast@image.ntua.gr

Dimitris Soufleros
Dsoufleros@gmail.com

Kostas Karpouzis
kkarpou@image.ntua.gr

Stefanos Kollias
National Technical University
of Athens
School of Electrical and
Computer Engineering
Image, Video and Multimedia
Systems Lab
9, Iroon Polytechniou str,
Athens, Greece
stefanos@cs.ntua.gr

ABSTRACT

We present a new dataset, ideal for Head Pose and Eye Gaze Estimation algorithm testings. Our dataset was recorded using a monocular system, and no information regarding camera or environment parameters is offered, making the dataset ideal to be tested with algorithms that do not utilize such information and do not require any specific equipment in terms of hardware.

Categories and Subject Descriptors

I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis - Tracking

General Terms

experimentation, algorithms

Keywords

Key words: Head Pose, Facial Feature Tracking, User Attention Estimation

1. INTRODUCTION

Estimation of head pose and eye gaze is a key issue for determining the focus of attention of a person. The algorithms existing in bibliography range from multi-camera [11] to monocular [10],[9], [3] systems and from systems that utilize specialized hardware [7] to simple methods in terms of equipment [10],[9], [3]. We present here a dataset created for algorithms relying solely on monocular systems, and exploiting no knowledge regarding camera intrinsic parameters. Our data is aimed at being tested with algorithms extracting head pose and eye gaze. Especially, eye gaze is a feature that, although thoroughly studied, is not accompanied by any official video database, and we present here a set of videos with this feature, both in conditions where the user's head is turned frontally, and not.

As for existing datasets in bibliography, a dataset offering similar information to the proposed one, is the one described in [12]. In that dataset, ground truth regarding head pose was offered by asking from the subjects to adjust a laser beam placed on their heads to different points ahead of them. Having the head fixed, they were asked to turn their eye gazes to specific positions. Thus, for a total of 20 people, 2220 static images were gathered, with different combinations of pose and gaze directionality. Main difference from the proposed dataset is that the background in [12] is controlled and the dataset is available in the form of static images, constituting itself not applicable for algorithms using tracking techniques.

More datasets, dealing with the problem of Head Pose estimation alone (thus, not including eye gaze) can be found in [8], where the data is offered, both in the form of video and static frames, but with no ground truth, in [5], where static images are offered, and in [4], where a set of grayscale images is offered, with facial feature landmarks included as well in ground truth. The Boston University dataset [2] offers a series of 45 videos from 5 subjects. Each video sequence consists of 200 frames and shows people moving freely their heads. Ground truth was acquired with the help of magnetic sensors.

2. DESCRIPTION OF THE DATASET

The proposed Head Pose and Eye Gaze dataset (HPEG dataset), created in lab conditions, consists of 20 color video sequences. Ten subjects -two female and eight male- participated and two videos were recorded for each (two sessions per person) of them. For video recording, we used a simple webcamera (Hercules Dualpix Exchange), that can record up to 1280×960 pixels frame size, at a rate of 30 fps. The software we used for recording the videos was the Hercules Webcam Station Evolution SE. No change of the camera parameters took place, in terms of white balance, saturation, luminance, contrast, with regards to the default values. All

AFFINE '09, November 6, 2009, Boston, MA, USA.

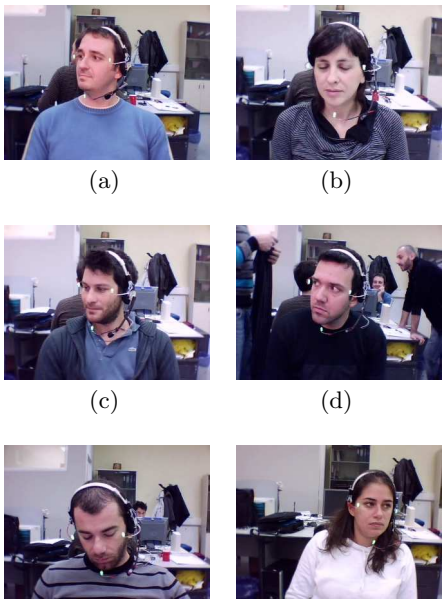

Figure 1: Examples from the session A

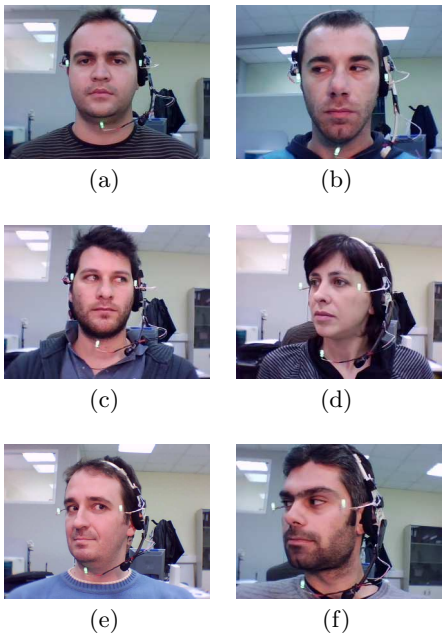

Figure 2: Examples from the session B

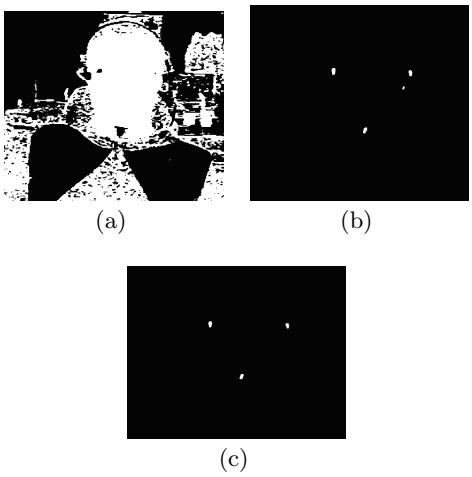

Figure 3: (a) Initial binary mask based on a* and b*;(b)Binary mask after logical AND between (a) and L map; (c)Final binary mask of LEDs after spatial information

videos were recorded at an analysis of 640×480 pixels, at 30fps, in uncompressed avi files. At naming the videos, the numbering means the subject ID and the letter (A or B) stands for the session.

During session A, the subject has to face the camera frontally, and, subsequently, he/she, is free to move towards any direction they want. Their movement can be, either translational, or rotational, or both at the same time. The camera is placed at a distance of almost one meter from the subject's head, almost opposite the eye level, while part of the torso is also visible. Also, the background covers a big part of the image, and it is a real workspace, with intense human action taking place at the background, at some of the videos; this constitutes our dataset more challenging in terms of algorithmic requirements as, many detectors/trackers might be "distracted" by human appearance in the background. The lighting conditions are the same as indoors conditions at an office environment and stay almost stable during the recordings. In total, every sequence of session A lasts 200 frames. Examples from the session A can be seen in Figure 1.

During session B, recorded under the same conditions as session A, the subject has to follow a pattern combining head and eye movements, as seen in figure 2. The camera is placed at a distance of almost 30cm from the subject. Session B was taken in order to create a dataset for eye gaze analysis, both in the case of frontally posed heads, and as a combination of head and eye movements. The protocol followed was the one seen in figure 2: At frame 2(a), the subject looks straight forward, both regarding his eyes, and his head. At frames 2(b) and 2(c), the subject still has his head rotated frontally, but his eyes look first towards one side (right or left), and then towards the other. At frame 2(d), the subject has rotated his head on his right, with the eyes looking frontally, aligned with the head directionality. In frames 2(e) and 2(f), the head keeps fixed, but the eyes turn left (looking at the camera) and right, respectively. The amount of frames in the second session varies from 349 to 430 per sequence.

3. EXTRACTION OF GROUND TRUTH

For the extraction of the ground truth, we used three LEDs, placed on the face. The aim was to track the LEDs positions at each frame, and extract information regarding Head Pose directionality. We used high intensity 15^o angle LEDs, and used a variable resistor to set the luminance at such levels that, a) they are visible under normal lighting conditions and b) they keep their color unchanged (it does not look white due to high luminosity). Furthermore, the limited angle of light of the LEDs has the disadvantage that, once placed on the head, and the head is rotated, they might not be visible, or cause our tracker to lose them. For this reason, the LEDs were placed vertically upwards and we wrapped around them a semi-transparent plastic case. By doing so, the LEDs shed their light in a homogenous way and could be visible, as long as they were not occluded by other objects. We chose green color for the light of the LEDs, as it does not mix with possible (usually red or blue) artifacts of a low quality webcamera, and differs significantly from the skin. Also, green is the less met color in an indoors environment or, at least, this was the case during the recordings.

Ground truth for Head Pose - sessions A and B - (yaw and pitch angles) was extracted based on the locations of the LEDs at every frame of the sequences. To this aim, we used a semi-automatic methodology, consisting in manually selecting the LEDs positions at the first frame of each sequence and tracking them in the subsequent frames. More specifically, the frames are converted to the corresponding $L^*a^*b^*$ colorspace and the means of the LEDs areas for the three channels are calculated. Similarly, the means of the three channels are also calculated for the rest part of each frame (*non*-LED areas). At subsequent frames, we compared each pixel's L^*,a^* and b^* values to those of the LED areas and non-LED areas. The two color ($a*$ and $b*$) channels were treated in a common coordinate system and pixels that appear to be closer to the mean values of LED than *non*-LED areas at the ($a*$,$b*$) subspace are declared as candidate LED pixels. Similarly, the same procedure was followed for luminance $L*$. The result was two maps that, when combined with the *AND* operator, gave a limited amount of candidate LED regions. Furthermore, information regarding the size and position of each LED in the previous frame is stored, and *non*-LED areas are excluded if they do not conform with these constraints. The above procedure is summarized in Figure 3. It should be noted that the LEDs were positioned on the subjects' faces in such a way that they would not become invisible at any time of the head rotation, by placing them a few centimeters ahead of the face level.

Having detected the LEDs positions at each frame, we used the equations described in [13] for determining the head rotation angles (*yaw* and *pitch*) at every frame, with regards to the camera plane, where the user is considered to be facing frontally when in parallel position with it. Here, instead of the eyebrows and the mouth, we used the positions of the three LEDs accordingly.

Eye gaze directionality (session B) is extracted with regards to the current head pose. For each position of the face, three classes of gaze directionality can be inferred: Looking straight forward, looking to the extreme left and to the extreme right. As these movements were controlled for all subjects in session B, ground truth in terms of class and time segments is offered by the dataset.

Table 1: Head Pose Estimation algorithm tested on the HPEG and BU datasets

	HPEG dataset	BU dataset
Yaw	5.41°	4.56°
Pitch	4.07°	3.82°

4. EXPERIMENTAL RESULTS

4.1 Head Pose Estimation

To test the effectiveness of using the HPEG dataset on testing a head pose estimation algorithm, we applied Distance Vector Field tracking, previously tested on the Boston University Dataset [2]. Distance Vector Fields have been used for detecting Facial features [1], and their usability can be extended to tracking eyes and mouth for inferring Head Pose directionality. More specifically, knowing the boundaries of the face skin region at every frame and, using them in relation to the positions of the facial features, permits the estimate of Head Pose orientation. An overview of the proposed algorithm is as follows: First, the face and facial features are detected, as in [1]. Subsequently, for tracking the face skin region, a skin area is extracted, between the eyes and the mouth and, based on its chrominance values and the face size, as calculated at the face detection step, a threshold is estimated. Pixels with chrominance values within certain limits, defined by the aforementioned threshold, with regards to the mean saturation value of the skin area, will be treated as Face pixels, thus, forming, a skin map. Morphological operators on the skin map further refine the result and the final detection is the skin region C_{skin} at each frame. To further increase the robustness of the skin tracking method, we utilized the rules defined in [6] and used an *AND* operator to combine the resulting map with C_{skin}.

Distance Vector Fields are image transformations assigning to every image pixel a vector pointing to its closest edge pixel. By applying a simple tracking procedure on DVFs, limited within the areas defined by C_{skin}, we achieved the estimate of features' positions during the video sequences, and further utilized their positions for Head Pose Estimation.

The eyes midpoint's horizontal distance H (as defined by the Distance Vector Fields' position) with regards to the skin midpoint, as defined by the tracked skin region, normalized by the inter-ocular distance at start-up, gives a good indicate of the horizontal head rotation (*yaw*). The same point's movement V with regards to the skin region lowermost point, also normalized with the inter-ocular distance, gives information with regards to the vertical head rotation (*pitch*). The above measurements, after being set to zero at the first frame, ($H = H - H_0$ and $V = V - V_0$, where 0 indicates their values at the first frame, where the user is supposed to be facing the camera frontally), are multiplied by factors f_1 and f_2 respectively, in order to be comparable to the ground truth information. Table 1 shows the results on session A of the HPEG dataset, for yaw and pitch, and shows the results obtained on the BU dataset. Results corresponding to our dataset are less accurate, probably due to the more challenging nature of the sequences in terms of video quality.

4.2 Eye Gaze Estimation

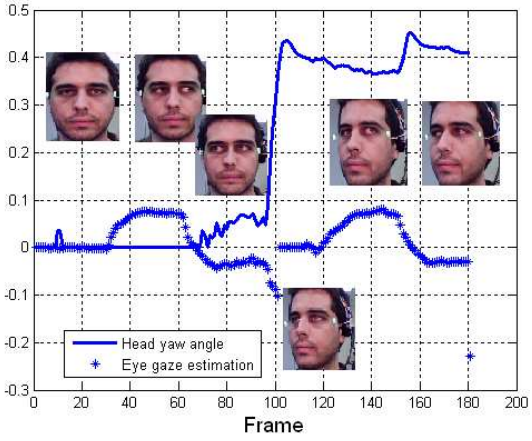

Figure 4: Example of sequence of session B and results on Eye Gaze Detection.

For testing eye gaze estimation algorithms, as mentioned earlier, our dataset provides with a set of ten image sequences of people posing both frontal and rotated head positions. The DVF tracking method, used in the previous subsection, was also employed here, and the iris centers were extracted from the eye areas [1]. Their positions were compared to the two lateral LEDs midpoint at every frame and any variations to their distances were decided to be due to eye gaze patterns. Figure 4 shows a typical example of gaze variation of a person, with the corresponding head poses. In the way described here, it is necessary to have knowledge of the time when the user is looking straight forward with his eyes and then, monitor the eye-gaze, setting it to zero when it is known that the person is looking straight forward. It can be seen from the figure that there is information regarding eye gaze variation. Future research shall consider training an automatic classifier for discriminating among gaze patterns using head orientation and eye gaze estimation. To this aim, the present dataset is expected to be of great support, as ground truth is already present for both features.

5. CONCLUSIONS

We presented a dataset appropriate for head pose and eye gaze estimation algorithm testings, in an uncontrolled environment with complex background. The dataset is accompanied by ground truth, both regarding head pose and eye gaze. The challenging part of our dataset is the inclusion of head pose and eye gaze in a common framework, which appear for the first time in video sequences in a public available dataset.

Acknowledgments

This work was partially funded by European Commission Projects 'METABO' (contract no. FP7-ICT-2007-216270) and by 'FEELIX Growing' (contract no. FP6 IST-045169).

6. REFERENCES

[1] S. Asteriadis, N. Nikolaidis, I. Pitas, and M. Pardàs. Detection of facial characteristics based on edge information. In *Second International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 247–252, Barcelona, Spain, 2007.

[2] M. L. Cascia, S. Sclaroff, and V. Athitsos. Fast, reliable head tracking under varying illumination: An approach based on robust registration of texture-mapped 3d models. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:322–336, 2000.

[3] C. Collet, A. Finkel, and R. Gherbi. Capre: a gaze tracking system in man-machine interaction. *JACIII*, 2(3):77–81, 1998.

[4] A. Georghiades, P. Belhumeur, and D. Kriegman. From few to many: Illumination cone models for face recognition under variable lighting and pose. *In IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643–660, 2001.

[5] N. Gourier, D. Hall, and J. L. Crowley. Estimating face orientation from robust detection of salient facial structures. *FG Net Workshop on Visual Observation of Deictic Gestures (POINTING)*, 2004.

[6] P. P. Jure Kovac and F. Solina. Human skin colour clustering for face detection. In *IEEE International Conference on Computer as a Tool*, volume 2, 2003.

[7] A. Kapoor and R. W. Picard. A real-time head nod and shake detector. In *in Proceedings from the Workshop on Perspective User Interfaces*, 2001.

[8] K. Messer, J. Kittler, M. Sadeghi, S. Marcel, C. Marcel, S. Bengio, F. Cardinaux, J. Czyz, S. Srisuk, M. Petrou, W. Kurutach, E. Kadyrov, B. Kepenekci, F. B. Tek, G. B. Akar, and F. Deravi. Face verification competition on the xm2vts database. In *fourth International Conference on Audio and Video Based Biometric Person Authentication*, pages 964–974, 2003.

[9] P. Smith, S. Member, M. Shah, and N. D. V. Lobo. Determining driver visual attention with one camera. *IEEE Trans. on Intelligent Transportation Systems*, 4:2003, 2003.

[10] J. Sung, T. Kanade, and D. Kim. Pose robust face tracking by combining active appearance models and cylinder head models. *International Journal of Computer Vision*, 80(2):260–274, 2008.

[11] M. Voit, K. Nickel, and R. Stiefelhagen. Multi-view head pose estimation using neural networks. In *Second Canadian Conference on Computer and Robot Vision (CRV)*, pages 347–352, Victoria, BC, Canada, 2005. IEEE Computer Society.

[12] U. Weidenbacher, G. Layher, P.-M. Strauss, and H. Neumann. A comprehensive head pose and gaze database. In *3rd IET International Conference on Intelligent Environments*, Ulm (Germany), Sept. 2007.

[13] A. Yilmaz and M. Shah. Automatic feature detection and pose recovery for faces. In *Proceedings of the Fifth Asian Conference on Computer Vision*, pages 284–289, 2002.