Robust Validation of Visual Focus of Attention using Adaptive Fusion of Head and Eye Gaze patterns

Stylianos Asteriadis[1], Kostas Karpouzis[1] and Stefanos Kollias[2]
National Technical University of Athens, Greece
Electrical and Computer Engineering Department
{stiast, kkarpou}@image.ntua.gr, stefanos@cs.ntua.gr

Abstract

We propose a framework for inferring the focus of attention of a person, utilizing information coming both from head rotation and eye gaze estimation. To this aim, we use fuzzy logic to estimate confidence on the gaze of a person towards a specific point, and results are compared to human annotation. For head pose we propose Bayesian modality fusion of both local and holistic information, while for eye gaze we propose a methodology that calculates eye gaze directionality, removing the influence of head rotation, using a simple camera. For local information, feature positions are used, while holistic information makes use of face region. Holistic information uses Convolutional Neural Networks which have been shown to be immune to small translations and distortions of test data. This is vital for an application in an unpretending environment, where background noise should be expected. The ability of the system to estimate focus of attention towards specific areas, for unknown users, is grounded at the end of the paper.

1. Introduction

Gaze directionality plays an important role since the very early years of our life: By watching their care takers' gaze directionality, small children learn to distinguish between important and less important events or objects [19]. Observing other people's focus of attention during meetings or social gatherings is a crucial factor for what is called *shared attention*[8]: People participating in events might be looking at something that has limited information to give, only because they want to declare their existence and attentiveness. A lot of works in recent bibliography have studied the issue of relating gaze with attention. In [23] the authors explore the correlation among one's own statements of attention towards an electronic material, the perception of others regarding his levels of attention and data from his gaze behavior. Gaze directionality has been the main feature for defining the levels of attentiveness of a person towards electronic material based on the amount of time a user spends looking at the object or, also, objects relative to the object of interest in [22]. More recent works regarding attention estimation on a multitude of targets are presented in [2] and [32]. In [2], the authors, using meeting events as context information, propose DBN modelling for inferring joint Visual Focus of Attention on a number of participants, in order to make assumptions regarding most probable targets that should attract more attention. The authors in [32] estimate focus of attention of participants in dynamic meeting scenarios, taking advantage of information coming from speech and motion activity of other participants.

In literature, various works exist for approaching the problem of gaze estimation, varying in terms of applications, cues and hardware set-up they use. A lot of work has been done for estimating the degree of concentration in driving conditions [7], [25], [9]. For example, in [9], the authors use stereoscopic techniques to estimate head and eye directionality, in order to simulate attentiveness in driving conditions. Similar, the problem of gaze estimation in conditions of Human-Agent interaction is under intense research [23, 22], [3], [18], [24]. However, these works are confined within certain bounds in terms of applicability or flexibility to use both head rotation and eye gaze as indicators, and suffer from the problem of multi-camera or intrusive systems.

In the proposed system, we describe a full-fledged methodology for estimating degrees of confidence of attention towards specific targets, using information coming from head rotation and eye gaze estimation in a common framework, with the usage of a single camera. Not a lot of works exist in bibliography regarding the issue of combining head pose and eye gaze, only with the usage of one ordinary camera, due, mainly, to the challenging nature of the problem. In the system described in [33], the authors use elastic graphs for the estimate of the horizontal head rotation. For eye gaze estimation, they employ gabor filters. Based on training data, they build lookup tables that match

the focus of attention with eye gaze and head pose calculations. Typical work on eye gaze and head pose estimation is the one described in [29], where the authors model heads with cylindrical models and, using the cylinder parameters, estimate the location of the eyes. These positions are projected on a normalized model view and are compared to reference positions in order to acquire eye gaze directionality. Head pose estimation (with no eye gaze estimates) is a more studied issue, however, and has been studied from many aspects, with works employing techniques based on holistic appearance [26], local information [17], [21], facial motion recovery [4], non-rigid models [6] and fusion of the above techniques for robust results [20].

2. System Overview

The proposed system uses Head Pose and Eye Gaze Estimation as inputs to a neural-fuzzy inference system which calculates the degree of visual attentiveness of a person towards specific areas. To this aim, a commonly used dataset [4] has been annotated regarding its participants' overall gaze towards the camera. In this work, we do not focus on inferring exact gaze estimation, but rather, we are interested in detecting degrees of confidence, through fuzzy logic, regarding hypotheses that a person is looking towards a specific point. A frame sequence dataset with clear annotation referring to focus of attention values, coming from both head rotation and eye gaze, to the authors' knowledge, is not publicly available and, thus, we chose to annotate accordingly a dataset taken under non-pretending conditions in terms of lighting and user movements, asking from annotators to declare the degree to which they think participants pay attention to the camera, positioned in front of them. Furthermore, the BU dataset, apart from the already available annotation regarding head pose, includes significant variation in terms of both head and eye gaze patterns, something that was expected - and, actually, desired - to be taken into account during annotation. Thus, although specific gaze annotation is not offered, the study we set up is a way to predict overall attentiveness towards the desired focus of attention. Future extensions of our work shall include more specialized datasets, developed in specific contexts (e.g. game-playing or e-learning environments).

For estimating Head Pose, both local and holistic information are used. Local information [1] consists in face and facial feature tracking, and is based on geometrical relations between facial features and face boundaries. Holistic information is extracted from the whole facial area and uses a set Convolutional Neural Classifiers [14], whose outputs are combined using linear regressors to estimate head pose. Fusion of both modalities aims at alleviating drawbacks of each of other: local techniques are highly dependent on correct tracking while holistic methods are sensitive to incorrect face segmentation. In this paper, we employ a Bayesian

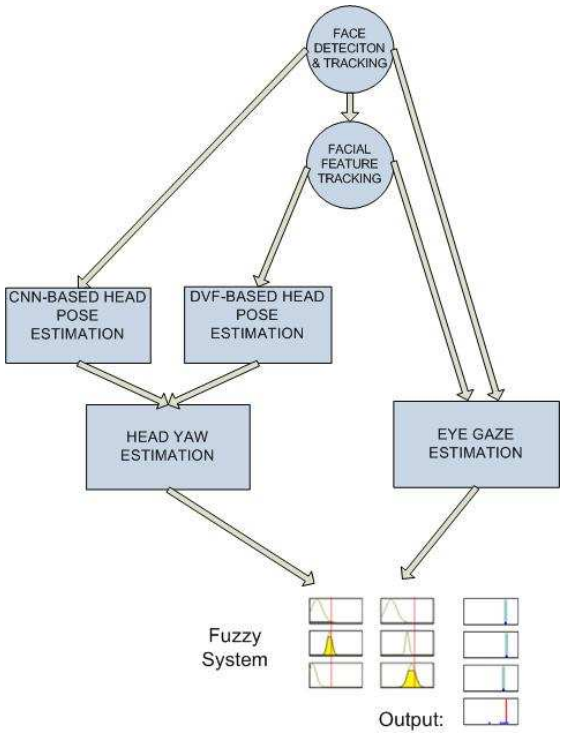

Figure 1. Overview of proposed methodology

Modality Fusion [28] scheme in order to model context information and reliability, taking into account necessary indicators.

In an attempt to estimate eye gaze, under head rotation, in this work, eye gaze estimation is achieved by using cylindrical models. An area around the eyes is modelled as a cylinder and it is distorted according to the head rotation (only rotations around the vertical axis are considered), towards the opposite direction. The positions of the eye centers in the new, rotated (yaw angle deprived) image are used to infer eye gaze directionality.

As the system is re-initialized when necessary, error accumulation is avoided, and the algorithm is able to adjust to scale changes, as all measurements are normalized with the inter-ocular distance as estimated at the initialization step. An overview of the proposed architecture is shown in Fig. 1.

The structure of the paper is as follows. In Section 3 the method for head pose estimation based on local features and holistic information is presented, and the proposed scheme for fusion is analyzed, while Section 4 explains how eye gaze information is extracted, even under head rotations. In Section 5 we discuss the methodology used for inference regarding the focus of visual attention of annotated data, as well as the experimental procedure we followed. Section 6 concludes the paper.

3. Feature tracking and Head Pose Estimation

A very important cue for estimating the degree of attention of a person towards a task he/she has in front of him/her is the rotation of his head. In bibliography, this problem is referred to (together with the estimate of its 3D world coordinates) as Head Pose Estimation. Here, we use a local technique based on DVFs [1] and we propose, as holistic information, a Convolutional Neural Networks [14] architecture and an inference scheme, adopting linear models between subclassifiers' outputs and head rotation space.

3.1. Feature tracking and Head Pose estimation using DVFs

Initially, the face and facial features are detected using the frontal Viola-Jones [31] face detection scheme followed by an ellipse fitting algorithm. Subsequently, the face region is tracked [1] based on each user's face chrominance model, learnt online, and his expected face size, as calculated at the face detection step. Learning personalized face chrominance models, in conjunction with expected facial area size constitute face tracking immune to color and lighting variations, as well as different scales at which a person can appear. Furthermore, these models can be learnt online and be re-trained each time the system re-initializes, so that error accumulation is avoided. The face bounding box is used so that facial characteristics are constrained within it, and provides input to the holistic technique described in the next section. The eyes and the mouth are detected and tracked using Distance Vector Fields (DVF), as was done in [1]. Distance Vector Fields assign a vector to every pixel of candidate feature areas, pointing to the closest edge pixel. In this way, each feature's geometry is encoded and tracking follows on this transform, instead of the image itself. Facial Feature tracking by comparing the Distance Vector Fields of successive images has given robust results for large and rapid head movements, while features' positions with regards to face boundaries have been used efficiently to estimate the *yaw* angle of the head [1], achieving accuracy on the Boston University dataset [4], equal to 4.4° (mean absolute error).

3.2. Head Pose estimation using CNNs

The architecture of a typical Convolutional Neural Network (CNN) is similar to that of a typical Neural Network, in the sense that it consists of layers of transformed versions of the input. More precisely, in a generic image recognition problem, an $N \times N$ image is used as input in the first layer. The image is convolved with a series of trainable filters of size $p \times p$ in the second layer ($C1$), resulting in feature maps of size $N - p + 1 \times N - p + 1$. In the third layer ($S2$), the feature maps are subsampled by a coefficient. The above characteristics of layers $C1$ and $S2$ guarantee the following: first, the ability of the network to learn robustly from a small amount of training data, within reasonable time, is feasible, as the number of the free parameters is significantly reduced, since feature maps' units share the same weights. Second, subsampling renders the network more resisting to small distortions or translations of the input image. Also, alternating convolutional and subsampling layers, makes it easy to form layers that start with detecting simple features (e.g edges, corners) and end up to combining features with each other in subsequent layers, to achieve information coming from spatial combination of them.

The ability of Convolutional Neural Networks for character recognition has been shown in [14] and it is enforced by the fact that they can learn, efficiently, spatial relations among characteristics, in contrast to typical Neural Networks. Furthermore, they do not require very precise alignment between training and test data, which is essential for situations where the user is moving freely in a scene with complex background. In this case, the boundaries of the face are usually not precisely tracked and the input to the Neural Network would not be exactly aligned with the training data.

In the proposed architecture (Fig. 3), training is done using stochastic Levenberg-Marquardt [15], and the hyperbolic tangent sigmoid function is used as activation function throughout the network. For training, the face dataset in [10] was used. The proposed inference scheme can be seen, schematically, in Fig. 2: Here, we have used the dataset in [10] to train 38 CNN classifiers of adjacent classes in the pose space, which was created using pairs of images centered to angles $\{-90^o, -45^o, 0, 45^o, 90^o\}$ of horizontal (yaw) rotation and $\{-60^o, 0, 60^o\}$ of vertical (pitch) rotation. The target values for training each CNN are the pairs $\{-1, 1\}$ or $\{1, -1\}$, depending on which class each training image belongs to. Face patches in training dataset were translated from 1 until 3 pixels towards all directions, and were mirrored around the vertical axis, in order to increase variability. Also, as face patches for testing originate from skin segmentation, there is usually a high degree of variability regarding the lower limits of the face (and neck) region. For this reason, all face patches for training and testing have been cropped in a way that face length is 1.3 times the face width. Furthermore, the effect of *roll* angle has been eliminated by rotating the face patch in a way that both eyes (see previous section) lay on the same horizontal level. The proposed CNN architecture can be seen in Fig. 3. Convolutional layers $C1$ and $C3$ consist of 6 feature maps, and layer $S2$ consists of 6 maps of dimensions half the size of $C1$. Layer $C4$ has 80 feature maps, while $F5$ consists of 10 neurons ending up to the output which gives two values within the range {-1,1}. For training, all images were normalized and were brought to 32×32 pixels. The convolutional layers use 7×7 kernels.

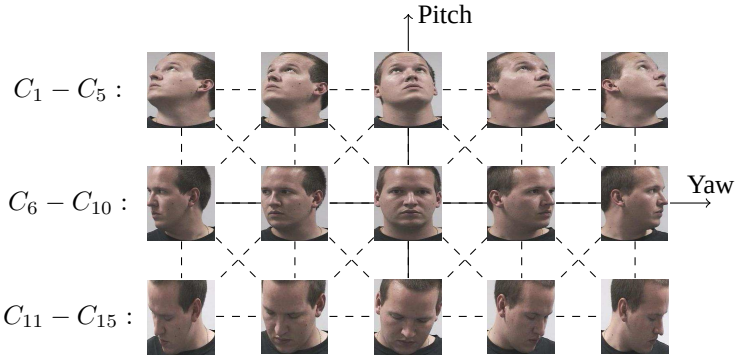

Figure 2. Head Pose classes used for training the Convolutional Neural Networks. Each trained CNN is denoted with a dashed line

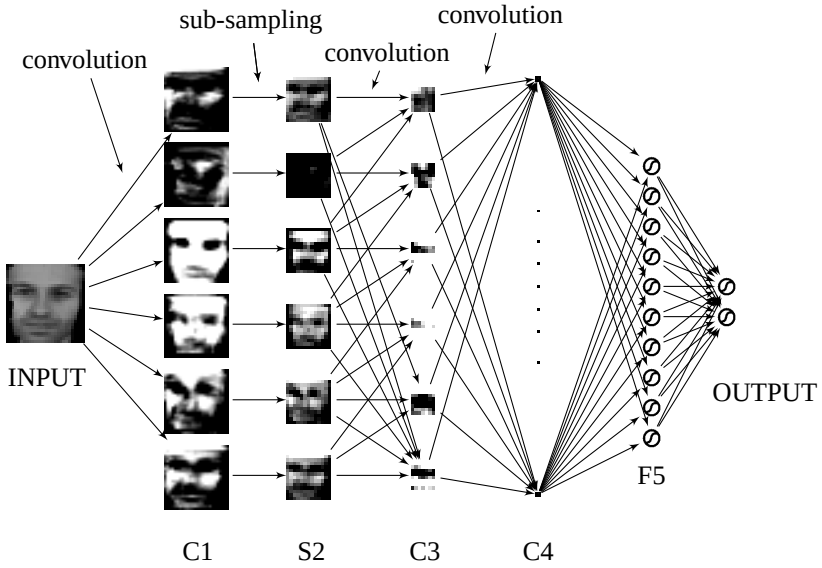

Figure 3. Convolutional Neural Network Architecture for Head Pose Estimation. $C1$, $S2$, $C3$ consist of 6 feature maps and $C4$ consists of 80, while $F5$ consists of 10 neurons

For inference, we propose a technique that uses information from the previous frame, by firing only those networks that included in their training data the class C_c whose center is closer to the *yaw* and *pitch* values of the previous frame. In this way, only a subset of CNNs are used at each frame, constituting the system faster and more reliable, as the possibility of erroneous classification is reduced. For the final estimate of the yaw (or pitch) angle, we adopt a linear relation between subclassifiers' output and actual difference from C_c. Thus, we propose regression models including the differences of CNN outputs, as well as the centre of class C_c. Considering the two outputs as $o1$ and $o2$, the difference $o1 - o2$ is a means of comparing class C_c against neighboring classes. Consequently, for each frame, we can have an overall output as a an 8-element vector consisting of such differences (o_{up}, o_{down}, o_{left}, o_{right}, $o_{up,left}$, $o_{up,right}$, $o_{down,left}$, $o_{down,right}$), with each element signifying the topological relation between the classes compared against C_c [1]. Horizontal (or vertical) rotation is calculated utilizing regression models on the above elements [2]

The method has been tested on the Boston University Face dataset and mean absolute error regarding *yaw* angle was $5.63°$(MAE).

3.3. Fusion of holistic with local approach using Bayesian Modality Fusion

Based on the observation that our local and holistic techniques have different levels of reliability, depending on the context of the interaction, in this paper, we take this param-

[1] If C_c is at the boundaries of the pose space, dummy classifiers giving output equal to 2 are hypothesized (setting o_1=1 and o_2=-1 for the existing C_c and non-existing class, respectively) for the missing hypothesized classes

[2] At estimating the horizontal rotation, o_{up} and o_{down} have been removed from the model. In a similar manner, o_{left} and o_{right} were omitted at estimating the pitch angle, thus, giving 6-element vectors (as well as the centre of class C_c) at the regression model

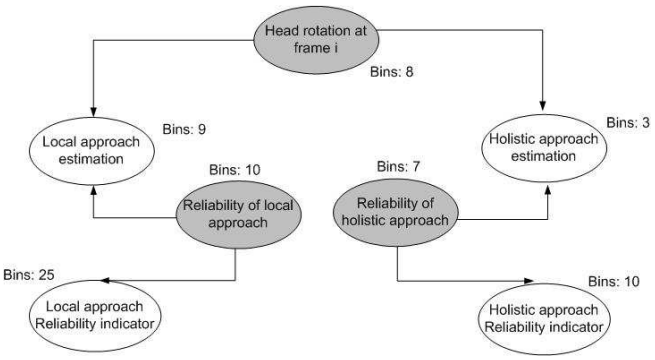

Figure 4. Bayesian Network Architecture used for fusing local with holistic technique.

eter into account during fusion. For this reason, we used Bayesian modality fusion, so that reliability of each cue is modelled, according to the phase of the interaction. The proposed architecture is explained in the next subsection.

3.3.1 Bayesian Networks for Head Rotation Estimation

In literature, the term *Bayesian Network* refers to a directional acyclic graph that represents the joint probability distribution for a set of random variables [11],[13]. In such networks, nodes are random variables and arcs stand for the statistical dependencies among pairs of nodes. Such dependencies, in a bayesian network model deterministic influences among the variables.

In this paper, estimated head horizontal rotation (yaw) has been considered to be a random, observable variable. On a second level, true head rotation affects visual systems' outputs (observable variables), which are also affected by each modality's reliability (hidden node). Reliability varies depending on the context of the interaction. As modality reliability cannot be observed during the sequence, an indirect way to infer it, is through measurable variables, correlated with it, namely modality reliability indicators. Figure 4 shows a schematic representation of the employed network, which is an adaptation of the scheme proposed in [28]. Graph nodes represent variables of interest, with the white ones corresponding to observable quantities and those with grey color corresponding to hidden variables. Node *Head rotation at frame* i is the final output variable (target node). The mean of the integral of the probability distribution of the target node gives the final estimate of head rotation.

3.3.2 Local information reliability indicator

As reliability indicator for local information, here is considered the fraction between vertical distance between mouth and eyes with eye distance:

$$rel_{DVF,y,i} = \frac{\parallel Eyes_{middle,i} - Mouth_{middle,i} \parallel}{\parallel Eyes_{right,i} - Eyes_{left,i} \parallel} \quad (1)$$

3.3.3 Holistic information reliability indicator

For each instance of the estimate of horizontal rotation with Convolutional Neural Networks, at frame i, $y_{CNN,i}$, the confidence value modelled as reliability indicator derives from equation 2:

$$rel_{CNN,y,i} = 1 - \frac{|y_{CNN,i} - m_{y,i:i-n+1}|}{std_{y,i:i-n+1}} \quad (2)$$

with $m_{y,i:i-n}$ and $std_{y,i:i-n}$ being average values and standard deviation for horizontal rotation for temporal windows of the n previous frames (here, we used n=5). The values of reliability indicators, under normal circumstances, are within specific values but, when the corresponding modality reliability is low, they can take arbitrary values [16].

3.3.4 Network parameters

Network training was based on learning conditional probability tables for the nodes which were learnt by quantizing variables into bins. The discretization that gave the optimum trade-off between variance and bias can be seen in Fig. 4. Tables parameters are learnt by simply counting (and normalizing) those frames where two events co-occur.

4. Eye Gaze Estimation

For estimating eye gaze, we propose a technique that models the face area around the eyes (Fig. 5) by a cylindrical shape, with pose parameters equal to $p = [\omega_\chi, \omega_y, \omega_z, t_\chi, t_y, t_z]$, where ω_χ, ω_y, ω_z the cylinder rotation angles and t_χ, t_y, t_z the translation parameters. As the input image is solely the area around the eyes, we considered t_χ and t_y to be equal to zero, while t_z is considered to be 80cm. Similar, ω_χ (pitch angle) is considered zero here, and ω_z (roll angle) is also considered to be null, since it can be eliminated by rotating the image, as the eye positions are known (the image is rotated so that both eyes lay on the same level). ω_y is the horizontal angle (yaw), as calculated in Section 3. Here, we considered that the camera focal length is f=500 (in pixels). Subsequently, the input image is warped so that ω_y is zero (Fig. 5). From the two new positions of the eye centers, the one used is that of the eye that is closer to the camera, as the error caused by perspective projection is smaller. Its position on the horizontal axis is then compared to that of a frame when the person is looking frontally, in order to estimate the gaze vector. The resulting value is normalized with the inter-ocular distance,

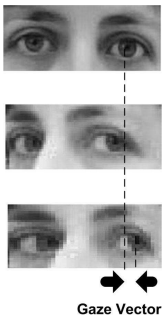

Figure 5. Extraction of Eye Gaze Vector. The eye position in the warped image (bottom) is compared to that of the frontal position, after yaw angle has been removed

as calculated at a frame when the person faces the camera frontally, in order to tackle scale variation.

5. Experimental procedure

5.1. Dataset Annotation

The Boston University dataset has been used here, and was annotated regarding the degree at which its participants are focused on the camera. For each sequence, we used 14 images, taken at intervals of 15 frames resulting to a total of 630 images (the dataset contains 45 sequences of 200 frames each). The extracted images were uploaded on a server and 102 people were asked to annotate up to 60 randomly selected images, each, regarding the degree of attention towards the camera they think the person in each image has (at a scale of 0-1, with 0 standing for complete distraction and 1 for gaze in the camera). In this way, each image has been annotated 8.75 times on average, and has been assigned the average of its annotations. Examples of images can be seen in Fig. 6. The use of the Boston University dataset, here, as our workbench, is due to the dataset's nature: the lighting conditions are normal and participants move freely, with high degree of spontaneity, changing both head rotations and their eye directionality. Thus, although the dataset offers ground truth regarding head pose only (by employing bayesian logic for fusion, the achieved overall head pose error on the Boston University dataset was 4.29° MAE/5.66° RMS, with the authors in [20] achieving an overall MAE error equal to 4.97° on the same dataset and the authors in [30] RMS error equal to 6.1° for horizontal head rotation), here, during the annotation set up for the current work, volunteers were expected to take into account eye gaze directionality as well for declaring their degree of confidence that someone is facing the camera or not.

5.2. Description of Inference System

Head pose and eye gaze are used as inputs to a Sugeno-type [27] fuzzy inference system to infer confidence values regarding focus of attention towards the camera, utilizing

(a) (b)

Figure 6. Examples of annotated images with annotations equal to 0.53 and 0.6, respectively.

the annotation described above, as ground truth data. Prior to training, our data were clustered using the sub-cluster algorithm described in [5]. This algorithm, instead of using a grid partition of the data, clusters them and, thus, leads to fuzzy systems deprived of the curse of dimensionality. For clustering, many radius values for the clusters were tried and the ones that gave the best trade-off between complexity and accuracy were 25° for head horizontal rotation, 0.15 for gaze vectors, and 0.40 for the output variable. The number of clusters created by the algorithm determines the optimum number of the fuzzy rules. After defining the fuzzy inference system architecture, its parameters (membership function centers and widths), are acquired by applying a least squares and back-propagation gradient descent method [12].

5.3. Experimental results on focus of attention estimation

Training of the bayesian network, as well as the Fuzzy Inference System was done by following a leave-one-out cross-validation method for each user, exempting all video sequences corresponding to him and using only those belonging to the rest of the participants. In this way, our system's aim is to be able to generalize and be used in applications where a user-specific calibration phase is supposed to be avoided. Taking into account that the overall settings of the dataset are non-pretending (every user moves in a personalized manner and lighting is normal), experimental performance shows that the system's ability to generalize to unknown users is promising. Testing for each user showed that the overall system was capable to estimate ground truth, as it was annotated by the users on the dataset of subsection 5.1, with an absolute error equal to 0.16.

To get a more precise picture of the system's ability to estimate those moments when the user is looking at specific points, raters' annotation, when larger than a certain threshold was considered to correspond to gaze patterns on the camera. When annotations were smaller than this threshold, it was considered that users were looking away from the camera. Visual inspection of the annotations and the corresponding images revealed that there was high variance when head would pose a high rotation with regards to the camera plane, but the eyes were actually looking at it. In such images, qualitative assessment of the annotation showed that raters would consider users looking at the cam-

era at a degree around ~ 0.5 out of 1 (see Fig. 6). Thus, setting a threshold at the fuzzy system's output, equal to T=0.5 for declaring a user as *looking at the camera*, overall recall and precision were 89% and 75%, respectively (f-measure=0.79).

In terms of speed, although the algorithm was developed using MATLAB, and no code optimization took place, on a Dual Core 2.26Ghz processor, DVF transform and feature tracking necessitate less than 100ms/frame, while a CNN subclassifier's output needs, on average, 10^{-4} sec to be estimated.

6. Conclusions and Future work

In this paper, the ability of a system to infer focus of attention of a user, towards a task in front of her or him, based on a combination of head pose and eye gaze directionality has been examined. For head pose estimation, we propose a hybrid technique, taking advantage of both holistic and local information, while eye gaze has been estimated using information from the same camera. Through experiments, it was shown that the estimates are reliable, indicating that the proposed methodology on head pose and eye gaze combination for a cumulative estimate of user focus of attention estimation is promising, especially taking into account that not a lot of work has been done on fusing these two cues in a non-calibrated, mono-camera environment. Current research is focusing on modelling user focus of attention on more than one areas or objects (in this paper, the camera), exploiting the possibilities of fuzzy logic towards that direction. Using raw instances of eye gaze and head pose, for estimating the exact gaze vector, and evaluation on appropriate frame sequences, is also within the aims of our current research. Our work is expected to support human-robot interaction environments, where the notions of shared attention and imitation are vital for natural dialogues, and adaptation to human preferences.

References

[1] S. Asteriadis, K. Karpouzis, and S. Kollias. Head pose estimation with one camera, in uncalibrated environments. In *Proceedings of the IUI2010, Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2010. 2, 3

[2] S. O. Ba and J.-M. Odobez. Multiperson visual focus of attention from head pose and meeting contextual cues. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(1):101–116, 2011. 1

[3] B. Brandherm, H. Prendinger, and M. Ishizuka. Interest estimation based on dynamic bayesian networks for visual attentive presentation agents. In *Proceedings of the 9th ACM International Conference on Multimodal Interfaces*, pages 346–349, 2007. 1

[4] 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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:322–336, 2000. 2, 3

[5] S. L. Chiu. Fuzzy Model Identification Based on Cluster Estimation. *Journal of Intelligent and Fuzzy Systems*, 2(3), 1994. 6

[6] T. Cootes, K. Walker, and C. Taylor. View-based active appearance models. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 227–232, 2000. 2

[7] T. D' Orazio, M. Leo, C. Guaragnella, and A. Distante. A visual approach for driver inattention detection. *Pattern Recognition*, 40(8):2341–2355, 2007. 1

[8] M. H. David, D. B. Grimes, A. P. Shon, and R. P. N. Rao. A probabilistic model of gaze imitation and shared attention. *Neural Networks*, 19(3):299–310, 2006. 1

[9] A. Doshi and M. Trivedi. Head and gaze dynamics in visual attention and context learning. In *IEEE CVPR Workshop on Visual and Contextual Learning*, 2009. 1

[10] N. Gourier, D. Hall, and J. Crowley. Estimating face orientation from robust detection of salient facial features. In *International Workshop on Visual Observation of Deictic Gestures (ICPR)*, Cambridge, UK, 2004. 3

[11] E. Horvitz, J. S. Breese, and M. Henrion. Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning*, 2(3):247–302, 1988. 5

[12] J.-S. R. Jang. ANFIS: Adaptive-Network-Based Fuzzy Inference System. *IEEE Transactions on Systems, Man, and Cybernetics*, 23:665–684, 1993. 6

[13] F. V. Jensen. *An Introduction to Bayesian Networks*. Springer-Verlag, 1996. 5

[14] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Handwritten digit recognition with a back-propagation network. In *Advances in Neural Information Processing Systems*, pages 396–404, 1990. 2, 3

[15] Y. LeCun, L. Bottou, G. Orr, and K. Muller. Efficient backprop. In G. Orr and M. K., editors, *Neural Networks: Tricks of the trade*. Springer, 1998. 3

[16] F. Liu, X. Lin, S. Z. Li, and Y. Shi. Multi-modal face tracking using bayesian network. In *Proceedings of IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003. 5

[17] B. Ma, S. Shan, X. Chen, and W. Gao. Head yaw estimation from asymmetry of facial appearance. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 38(6):1501–1512, 2008. 2

[18] Y. Matsumoto, T. Ogasawara, and A. Zelinsky. Behavior recognition based on head pose and gaze direction measurement. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, 2000. 1

[19] M. Morales, P. Mundy, C. delgado, M. Yale, R. Neal, and H. Schwartx. Gaze following, temperament, and language development in 6-month-olds: A replica and extension. *Infant Behavior and Development*, 23(2):231–236, 2000. 1

[20] L.-P. Morency, J. Whitehill, and J. Movellan. Generalized adaptive view-based appearance model: Integrated framework for monocular head pose estimation. In *Proceedings*

of the IEEE International Conference on Face and Gesture Recognition, 2008. 2, 6

[21] M. H. Nguyen, J. Perez, and F. D. la Torre. Facial feature detection with optimal pixel reduction svm. In *Proceedings of the 8th IEEE International Conference on Automatic Face and Gesture Recognition*, pages 1–6, 2008. 2

[22] P. Qvarfordt and S. Zhai. Conversing with the user based on eye-gaze patterns. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, page 230. ACM, 2005. 1

[23] R. Ishii and I. Y. Nakano . An empirical study of eye-gaze behaviors: Towards the estimation of conversational engagement in human-agent communication. In *Proceedings of the IUI2010, Workshop on Eye Gaze in Intelligent Human Machine Interaction*, 2010. 1

[24] C. Sidner, C. Lee, C. Kidd, N. Lesh, and C. Rich. Explorations in engagement for humans and robots. *Artificial Intelligence*, 166(1-2):140–164, 2005. 1

[25] P. Smith, M. Shah, and N. da Vitoria Lobo. Determining driver visual attention with one camera. *IEEE transactions on intelligent transportation systems*, 4(4):205–218, 2003. 1

[26] R. Stiefelhagen. Estimating Head Pose with Neural Networks - Results on the Pointing04 ICPR Workshop Evaluation Data. In *Pointing 04 Workshop (ICPR)*, Cambridge, UK, 2004. 2

[27] T. Takagi and M. Sugeno. Fuzzy identification of systems and its applications to modelling and control. *IEEE Trans. Syst Man Cybern*, 15(1):116–132, 1985. 6

[28] K. Toyama and E. Horvitz. Bayesian modality fusion: Probabilistic integration of multiple vision algorithms for head tracking. In *4th Asian Conference on Computer Vision (ACCV)*, 2000. 2, 5

[29] R. Valenti, N. Sebe, and T. Gevers. Visual gaze estimation by joint head and eye information. In *International Conference on Pattern Recognition*, 2010. 2

[30] R. Valenti, Z. Yucel, and T. Gevers. Robustifying eye center localization by head pose cues. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 6

[31] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, volume 1, pages 511–518, 2001. 3

[32] M. Voit and R. Stiefelhagen. 3d user-perspective, voxel-based estimation of visual focus of attention in dynamic meeting scenarios. In *ICMI-MLMI*, page 51, 2010. 1

[33] U. Weidenbacher, G. Layher, P. Bayerl, and H. Neumann. Detection of head pose and gaze direction for human-computer interaction. In *Perception and Interactive Technologies*, pages 9–19, 2006. 1