Memoirs of the Scientific Sections of the Romanian Academy
Tome XXXIX, 2016

COMPUTER SCIENCE

A FEEDBACK CONTROL, GAZE FOLLOWING APPROACH FOR HUMAN-ROBOT INTERACTION

SORIN M. GRIGORESCU and FLORIN MOLDOVEANU

Department of Automation
Transilvania University of Brasov
5, Mihai Viteazu street, 500174 - Brasov, Romania
Corresponding author: s.grigorescu@unitbv.ro

Gaze following is one of the key requirements for successful *Human-Robot Interaction* (HRI). In this paper, a feedback controlled approach for facial features detection and gaze following is proposed. The goal of the method is to cope with uncertainties present in the context of HRI, namely different poses, occlusion and variable illumination conditions. The key elements of the method are the local and spatial estimators of the facial features, the *Gaussian Mixture Model* (GMM) used for segmenting the face, as well as the feedback control way in which the parameters of the whole processing chain are adapted. The system has been evaluated against temporal sequences of moving human agents, acquired *via* a stereo imaging system mounted on a mobile robotic platform. As performance metrics, the mean and the maximal normalized deviations between the manually determined ground truth and the estimated positions of the facial features have been chosen.

Keywords: gaze estimation, feedback control in image processing, facial features detection, Human-Robot Interaction.

1. INTRODUCTION

Head movements are commonly interpreted as a vehicle of interpersonal communication. For example, in daily life, human beings observe head movements as the expression of agreement or disagreement in a conversation, or even as a sign of confusion. On the other hand, gaze shifts are usually an indication of intent, as they commonly precede action by redirecting the sensorimotor resources to be used. As a consequence, sudden changes in gaze direction can express alarm or surprise. Gaze direction can also be used for directing a person to observe a specific location. To this end, during their infancy, humans develop the social skill of *joint attention*, which is the means by which an agent looks at where its interlocutor is looking at by producing an eye-head movement that attempts to yield the same focus of attention.

As artificial cognitive systems with social capabilities become more and more important due to the recent evolution of robotics towards applications where complex and human-like interactions are needed, basic social behaviors, such as joint attention, have increasingly become important research topics in this field. Figure 1 illustrates the ROVIS[1] (*Robust Vision and Control Laboratory*) gaze following

[1] http://rovis.unitbv.ro

system at work, in the context of joint attention for *Human Robotic Interaction* (HRI). Gaze following thus represents an important part of building a social bridge between humans and computers. Researchers in robotics and artificial intelligence have been attempting at accurately reproducing this type of interaction in the last couple of decades and, although much progress has been made [8], dealing with perceptual uncertainty still renders it difficult for these solutions to work adaptively.

Fig. 1. Gaze following in the context of joint attention for HRI, using the ROVIS system on a Neobotix MP 500 mobile platform.

Gaze following is an example for which the performance of artificial systems is still far from human adaptivity. In fact, the gaze following adaptivity problem can be stated as follows: how can gaze following be implemented under non-ideal circumstances (perceptual uncertainty, incomplete data, dynamic scenes, etc.)?

Feature detection represents a subtopic within the head pose estimation problem. Accurate estimate for the eye, nose or mouth represents an intermediate stage, in which essential information used by the geometrical approach for head pose estimation is computed. Mouth recognition is dealt with methods such as the ones suggested in [6] and [9]. Both methods use a ROI extracted after head segmentation, in which the mouth is approximately segmented, after a color space conversion is performed (such as RGB to HSI (*Hue, Saturation, Intensity*) [6], or RGB to Lab [9]). On the other hand, nose detection algorithms use Boosting classifiers, commonly trained with Haar-like features [1], or 3D information of the face, as in [13]. The shape-based algorithm proposed in [11], built on the isophote curvature concept, *i.e.* the curve that connects points of the same intensity, is able to deliver accurate eye localization from a web camera. Eye location can be determined using a combination of Haar features [12], dual orientation Gabor filters and eye templates, as described in [5].

In the following, we propose a robust solution to facial feature detection for human-robot interaction based on a feedback control approach implemented at image processing level, for the automatic adaptation of facial features detection system's parameters. The goal is to obtain a real-time gaze following estimator capable

of dealing with perceptual uncertainty and incomplete data. The expected outcome of this project will be an autonomous system, with the ability of robustly estimating gaze's direction of interlocutors within the context of joint attention in HRI.

2. FEEDBACK CONTROL IN IMAGE PROCESSING

In industrial or real world robotic applications, the purpose of the image processing system is to understand the surrounding environment of the robot through visual information.

Low level image processing deals with pixel wise operations aiming at improving the input images and also at separating the objects of interest from the background. Both inputs and outputs of low level blocks are images. The second type modules, which deal with high level visual information, are connected to low level operations through the feature extraction component which converts the input images to abstract data, describing the imaged objects of interest. For the rest of the high level operations, both the inputs and outputs are abstract data. The importance of the quality of the results coming from low level stages is related to the requirements of high level image processing [3]. Namely, in order to obtain a proper visual understanding of the imaged environment at high level stage, the inputs coming from low level have to be reliable.

The sequential, feedback free approach has an impact on the final perception result, since each operation in the chain is applied sequentially, with no information between the different levels of processing. In other words, low level image processing is done regardless of the requirements of higher levels. For example, if the segmentation module fails to provide a good output, all subsequent steps will fail. In [7] and [2], the inclusion of feedback structures within vision algorithms for improving the overall robustness of the chain is suggested. In the proposed approach, the parameters of low level image processing are adapted in a closed-loop manner in order to provide reliable input data to higher levels of processing.

The basic diagram, from which the feedback mechanisms for machine vision are derived in this paper, is plotted in Figure 2. In such a control system, the control signal u, or *actuator variable*, is an image processing parameter, whereas the *controlled variable* y is a measure of feature extraction quality.

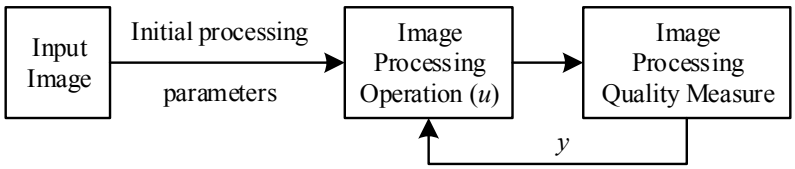

Fig. 2. Feedback adaptation of an image processing operation.
The image processing quality measure y is used as a feedback control variable for adapting the parameters of the vision algorithms using actuator u.

3. ROBUST GAZE FOLLOWING

The gaze following image processing chain, depicted in Figure 3, contains four main steps. We assume that the input is an 8-bit gray-scale image $I = J^{V \times W}$, of width V and height W, containing a face viewed either from a frontal or profile direction, where $J = \{0,...,255\}$. Symbol (v, w) represents the 2D coordinates of a specific pixel. The face region is obtained from a face detector.

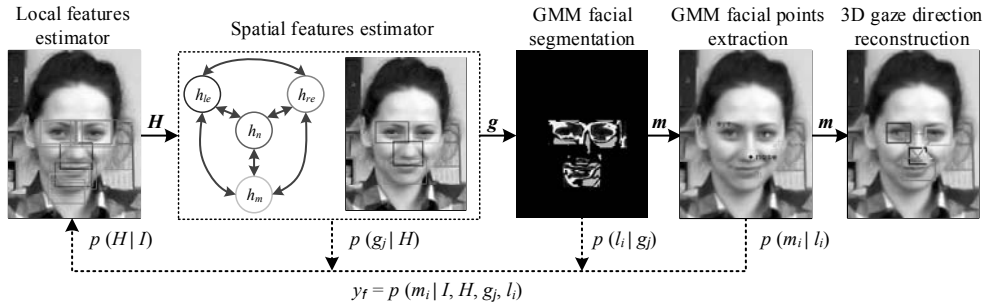

Fig. 3. Block diagram of the proposed gaze following system for facial feature extraction and 3D gaze orientation reconstruction. Each processing block within the cascade provides a measure of feature extraction quality, fused within the controlled variable y_f (see Eq. 2).

Firstly, a set of facial features ROI hypotheses $H \in \{h_{le}, h_{re}, h_n, h_m\}$, consisting of possible instances of the left h_{le} and right h_{re} eyes, nose h_n and mouth h_m, are extracted using a local features estimator which determines the probability measure $p(H \mid I)$ of finding one of the searched local facial region. The number of computed ROI hypotheses is governed by a probability threshold T_h, which rejects hypotheses with a low $p(H \mid I)$ confidence measure. The choice of the T_h threshold is not a trivial task when considering time critical systems, such as the gaze estimator which, for a successful HRI, has to deliver the 3D gaze orientation of the human subject in real-time. The lower T_h is, the higher the computation time. On the other hand, an increased value for T_h would reject possible "true positive" facial regions, thus leading to a failure in gaze estimation. As explained in the following, to obtain a robust value for the hypotheses selection threshold, we have chosen to adapt T_h with respect to the confidence values provided by the subsequent estimators from Figure 3, which take as input the facial regions hypotheses. The output probabilities coming from these estimation techniques, namely the spatial estimator and the GMM for pointwise feature extraction, are used in a feedback manner within the extremum seeking control paradigm.

Once the hypotheses vector H has been build, the facial features are combined into the spatial hypotheses $g = g_0, g_1, ..., g_n$, thus forming different facial regions

combinations. Since one of the main objective of the presented algorithm is to identify facial points of frontal, as well as of profile faces, a spatial vector s_i is composed from either four or three facial ROIs:

$$g_i = \{h_0, h_1, h_2, h_3\} \cap \{h_0, h_1, h_2\} \tag{1}$$

where $h_i = \{h_0, h_1, h_2, h_3\}$.

Extraction of the best spatial features combination can be seen as a graph search problem $g_j = f : G(g, E) \to \Re$, where E are the edges of the graph connecting the hypotheses in g.

4. PERFORMANCE EVALUATION

4.1. EXPERIMENTAL SETUP

In order to test the performance of the proposed gaze following system, the following experimental setup has been prepared.

The system has been evaluated on the *Labeled Faces in the Wild* (LFW) database [4]. LFW consists of 13,233 images, each with a size of $250 \times 250\ px$. In addition to the LFW database, the system has been evaluated on an Adept Pioneer 3-DX mobile robot equipped with RGB-D sensor delivering $640 \times 480\ px$ size color and depth images. The goal of the scenarios is to track the facial features of the human subject in the HRI context. The error between the real and estimated facial feature's locations was computed offline.

For evaluation purposes, two metrics have been used:

- the mean normalized deviation between the ground truth and the estimated positions of the facial features:

$$d(m, \hat{m}) = \tau(m) \frac{1}{k} \sum_{i=0}^{k-1} \left\| m_i - \hat{m}_i \right\|, \tag{2}$$

where k is the number of facial features, m and $\hat{m}$ are the manually and estimated annotated positions of the eyes, nose and mouth, respectively, and $\tau(m)$ is a normalization constant:

$$\tau(m) = \frac{1}{\left\| (m_{le} + m_{re}) - m_m \right\|} \tag{3}$$

- the maximal normalized deviation:

$$d^{\max}(m, \hat{m}) = \tau(m) \max_{j=0,\ldots,k-1} \left\| m_i - \hat{m}_i \right\| \tag{4}$$

4.2. COMPETING DETECTORS

The proposed gaze following system has been tested against three open-source detectors.

1) *Independent facial feature extraction*: The detector, based on the Viola-Jones boosting cascades, returns the best detected facial features, independent on their spatial relation. The point features have been considered to be the centers of the computed ROIs.

The boosting cascades, one for each facial feature, have been trained using a few hundred samples for each eye, nose and mouth. Searching has been performed several times at different scales, with Haar-like features used as inputs to the basic classifiers within the cascade. From the available ROI hypotheses, the one having the maximum confidence value has been selected as the final facial feature.

2) *Active Shape Models*: An *Active Shape Model* (ASM) estimates a dense set of feature points distributed around face contours such as eyes, nose, mouth, eyebrows, or chin. An ASM is initially trained using a set of manually marked contour points.

The open-source AsmLib, based on OpenCV, has been used as candidate detector. ASM is trained from manually drawn face contours. The trained ASM model calculates the main variations in the training dataset using *Principal Component Analysis* (PCA), which enables the model to automatically recognize if a contour is a face contour. PCA is used to find the mean shape and main variations of the training data with respect to mean shape. After finding the shape model, all training objects are deformed to the main shape, and the pixels are converted to vectors. The positions of the contours at each search step are corrected by the usage of lines perpendicular to the control points of the contour. After creating the ASM model, an initial contour is deformed by finding the best texture match for the control points. This is an iterative process, in which the movement of the control points is limited by what the ASM model recognizes from the training data as a "normal" face contour.

3) *Flandmark*: *Flandmark* [10] is a deformable part model detector of facial features, where detection of the point features is treated as an instance of structured output classification. The algorithm is based on a *Structured Output Support Vector Machine* (SO-SVM) classifier for a supervised learning of the parameters for facial points detection from examples. The objective function of the learning algorithm is directly related to the performance of the resulting detector, which is controlled by a user-defined loss function.

Comparatively with our gaze following system, which uses a segmentation step for determining the pointwise location of facial features, Flandmark considers the centers of the detected ROIs as the point location of the eyes, nose, and mouth.

The mean and maximal deviation metrics were used to compare the accuracy of the four tested detectors with respect to the ground truth values available from benchmark databases. Especially for the evaluation of the computation time, the algorithm has also been tested on a mobile robotic platform.

The cumulative histograms of the mean and maximal normalized deviation are shown in Figure 4 for frontal and profile faces. In all cases, the proposed estimator delivered an accuracy value superior to the ones given by the competing detectors. If the accuracy difference between our algorithm and that of Flandmark is relatively low for the frontal faces, it actually increases when the person's face is imaged from a profile view.

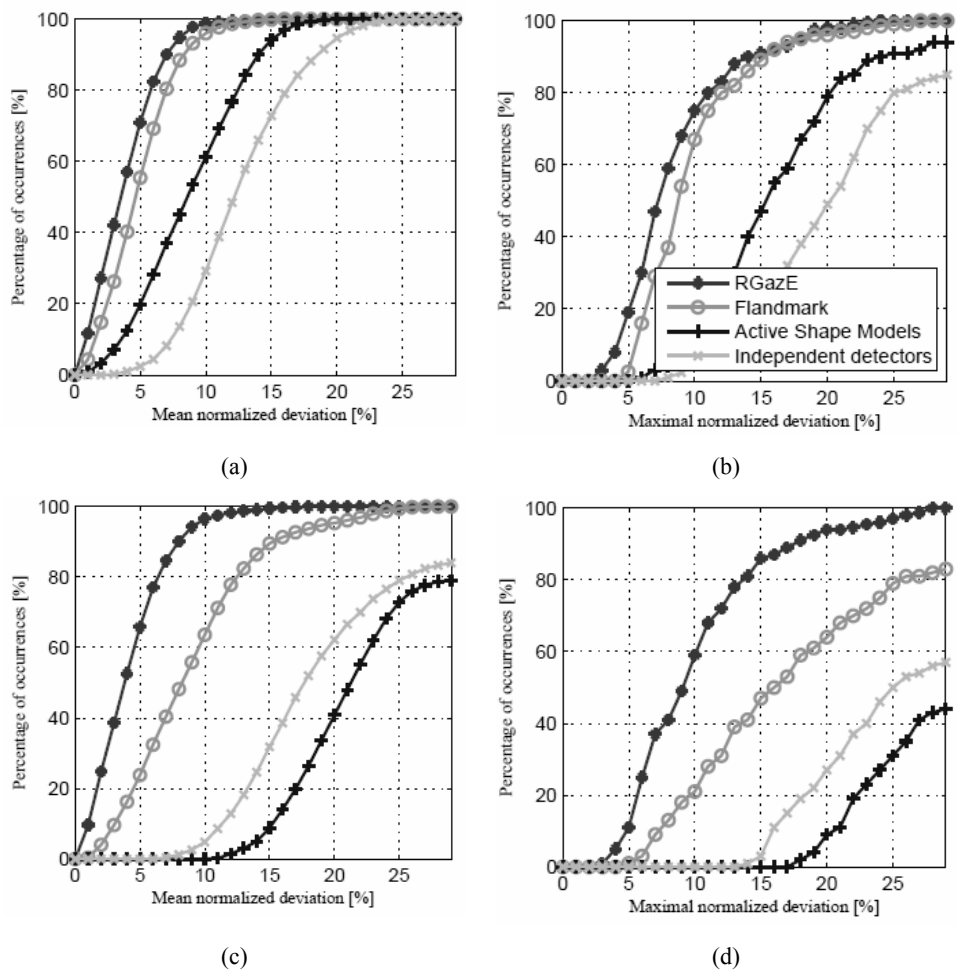

Fig. 4. Cumulative histograms for the mean and maximal normalized deviation shown for all competing detectors applied on video sequences with frontal (a, b) and profile (c, d) faces.

An interesting observation can be made when comparing the independent detectors with the ASM one. Although the ASM outperforms independent facial feature extraction on frontal faces, it does not perform well when the human

subjects are viewed from the lateral side. This is due to the training nature of the ASM, whose input training data is made of points spread over the whole frontal area (*e.g*. eyes, eyebrows, nose, chin, cheeks, etc.).

5. CONCLUSIONS

In this paper, a robust facial features detector for 3D gaze orientation estimation has been proposed. The solution is able to return a reliable gaze estimate, even if only a partial set of features is available, with a clear indication of the uncertainty involved. The paper brings together algorithms for facial feature detection, machine learning and control theory. During the experiments, we have investigated the system response and compared the results to ground truth values. As shown in the experimental results section, the method performed well with respect to various testing scenarios. As future work, the authors consider the possibility of extending the framework for a simultaneous gaze estimation of multiple interlocutors and adaptation of the algorithm with respect to robot's egomotion.

Acknowledgement. We hereby aknowledge the structural founds project PRO-DD (POS-CCE, O.2.2.1., ID 123, SMIS 2637, ctr. No 11/2009) for providing the infrastructure used in this work.

Authors' contribution: Both authors had equal contribution in writing this paper.

REFERENCES

1. GONZALEZ-ORTEGA D., DIAZ-PERNAS F., MARTINEZ-ZARZUELA M., ANTON-RODRIGUEZ M., DIEZ-HIGUERA J., BOTO-GIRALDA D., *Real-time nose detection and tracking based on ADABoost and optical flow algorithms*, Intelligent Data Engineering and Automated Learning. Springer, Berlin, 2009, **5788**, 142–150.
2. GRIGORESCU S.M., *Robust machine vision for service robotics*, Ph.D. dissertation, Bremen University, Institute of Automation, Bremen, Germany, June 2010.
3. HOTZ L., NEUMANN B., TERZIC K., *High-level expectations for low-level image processing*, KI 2008: Advances in Artificial Intelligence. Springer-Verlag Berlin Heidelberg, 2008.
4. HUANG G.B., RAMESH M., BERG T., LEARNED-MILLER E., *Labeled faces in the wild: A database for studying face recognition in unconstrained environments*, University of Massachusetts, Amherst, Tech. Rep. 07–49, October 2007.
5. KE, L., KANG J., *Eye location method based on Haar features*, 2010 3rd International Congress on Image and Signal Processing, 2010, **2**, 92 –929.
6. PANTIC M., TOMC M., ROTHKRANTZ L., *A hybrid approach to mouth features detection*, 2001 IEEE International Conference on Systems, Man, and Cybernetics, 2001, **2**, 1188–1193.
7. RISTIC D., *Feedback structures in image processing*, Ph.D. dissertation, Bremen University, Institute of Automation, Bremen, Germany, Apr. 2007.
8. SCASSELLATI B., *Theory of mind for a humanoid robot*, Autonomous Robots, 2002, **12** (1999), 13–24.
9. SKODRAS E., FAKOTAKIS N., *An unconstrained method for lip detection in color images*, 2011 IEEE International Conference on Acoustics, Speech and Signal Processing, 2011, 1013–1016.

10. URICAR M., FRANC V., HLAVAC V., *Detector of facial landmarks learned by the structured output SVM*, VISAPP '12: Proceedings 7th International Conference on Computer Vision Theory and Applications, G. Csurka and J. Braz, Eds., **1**, Portugal: SciTePress — Science and Technology Publications, February 2012, 547–556.
11. VALENTI R., SEBE N., GEVERS T., *Combining head pose and eye location information for gaze estimation*, IEEE Transaction on Image Processing, 2011.
12. VIOLA P., JONES M., *Rapid object detection using a boosted cascade of simple features*, in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 2001.
13. WERGHI N., BOUKADIA H., MEGUEBLI Y., BHASKAR H., *Nose detection and face extraction from 3D raw facial surface based on mesh quality assessment*, 36th Annual Conference on IEEE Industrial Electronics Society, 2010, 1161–1166.

Received February 11, 2016