

Human–Robot Interaction Through Robust Gaze Following

Sorin M. Grigorescu and Gigel Macesanu

1 Abstract In this paper, a probabilistic solution for gaze following in the context
2 of joint attention will be presented. Gaze following, in the sense of continuously
3 measuring (with a greater or a lesser degree of anticipation) the head pose and gaze
4 direction of an interlocutor so as to determine his/her focus of attention, is impor-
5 tant in several important areas of computer vision applications, such as the devel-
6 opment of nonintrusive gaze-tracking equipment for psychophysical experiments in
7 Neuroscience, specialized telecommunication devices, *Human–Computer Interfaces*
8 (HCI) and artificial cognitive systems for *Human–Robot Interaction* (HRI). We have
9 developed a probabilistic solution that inherently deals with sensor models uncer-
10 tainties and incomplete data. This solution comprises a hierarchical formulation of
11 a set of detection classifiers that loosely follows how geometrical cues provided by
12 facial features are used by the human perceptual system for gaze estimation. A quan-
13 titative analysis of the proposed architectures performance was undertaken through a
14 set of experimental sessions. In these sessions, temporal sequences of moving human
15 agents fixating a well-known point in space were grabbed by the stereovision setup
16 of a robotic perception system, and then processed by the framework.

AQI

17 1 Introduction

18 Head movements are commonly interpreted as a vehicle of interpersonal commu-
19 nication. For example, in daily life, human beings observe head movements as an
20 expression of agreement or disagreement in a conversation, or even as a sign of con-
21 fusion. On the other hand, gaze shifts are usually an indication of intent, as they
22 commonly precede action by redirecting the sensorimotor resources to be used. As a

S.M. Grigorescu (✉) · G. Macesanu
Department of Automation, Transilvania University of Brasov, Mihai Viteazu 5,
500174 Brașov, Romania
e-mail: s.grigorescu@unitbv.ro

G. Macesanu
e-mail: gigel.macesanu@unitbv.ro

© Springer International Publishing Switzerland 2017
P. Kulczycki et al. (eds.), *Information Technology and Computational Physics*,
Advances in Intelligent Systems and Computing 462,
DOI 10.1007/978-3-319-44260-0_10

165



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

23 consequence, sudden changes in gaze direction can express alarm or surprise. Gaze
 24 direction can also be used for directing a person to observe a specific location. To this
 25 end, during their infancy, humans develop the social skill of *joint attention*, which is
 26 the means by which an agent looks at where its interlocutor is looking at by producing
 27 an eye-head movement that attempts to yield the same focus of attention. Over nine
 28 months of age, infants are known to begin to engage with their parents/caregivers in
 29 an activity in which both look at the same target through joint attention.

30 As artificial cognitive systems with social capabilities become more and more
 31 important due to the recent evolution of robotics towards applications where complex
 32 and human-like interactions are needed, basic social behaviors such as joint attention
 33 have increasingly become important research topics in this field. Figure 1 illustrates
 34 the ROVIS¹ (*Robust Vision and Control Laboratory*) gaze following system at work,
 35 under the context of joint attention for *Human Robotic Interaction* (HRI). Gaze fol-
 36 lowing thus represents an important part of building a social bridge between humans
 37 and computers. Researchers in robotics and artificial intelligence have been attempt-
 38 ing to accurately reproduce this type of interaction in the last couple of decades, and,
 39 although much progress has been made [1], dealing with perceptual uncertainty still
 40 renders it difficult for these solutions to work adaptively.

41 Gaze following is an example for which the performance of artificial systems is
 42 still far from human adaptivity. In fact, the gaze following adaptivity problem can
 43 be stated as follows: how can gaze following be implemented under nonideal cir-
 44 cumstances (perceptual uncertainty, incomplete data, dynamic scenes, etc.)? Figure 2
 45 demonstrates how incomplete data, arguably the issue where the lack of adaptivity
 46 and underperformance of artificial systems are most apparent, might influence the
 47 outcome of gaze following.

48 In the following text, we propose a robust solution to facial feature detection for
 49 human–robot interaction based on (i) a feedback control system implemented at the
 50 image processing level for the automatic adaptation of the system’s parameters, (ii) a

¹<http://rovis.unitbv.ro>.

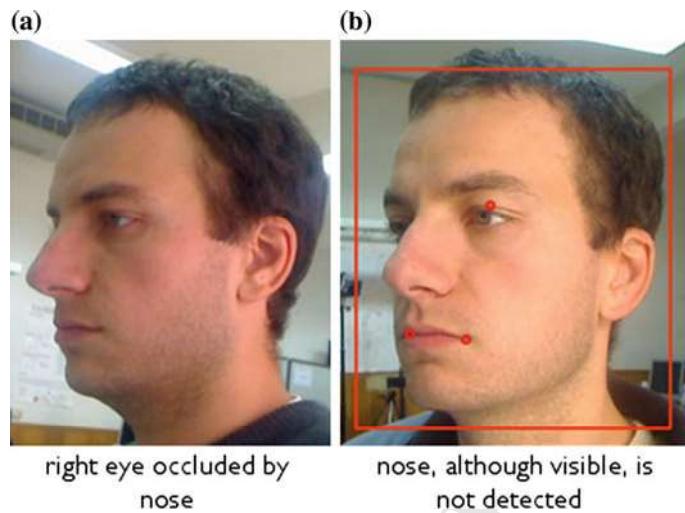


Fig. 2 Examples of probable gaze following failure scenarios due to incomplete data: facial features occluded in profile views (a), or failure of feature detection algorithms (b)

Editor Proof

51 cascade of facial features classifiers, and (iii) a *Gaussian Mixture Model* (GMM) for
 52 facial points segmentation. The goal is to obtain a real-time gaze following estimator
 53 which can cope with uncertainties and incomplete data. The proposed system aims at
 54 the robust computation of the human gaze direction in the context of joint attention
 55 for HRI.

56 2 Related Work

57 2.1 Gaze Following

58 In recent years, the problem of gaze following has been extensively studied. Physiological
 59 investigations have demonstrated that the brain estimates the gaze as a mixture
 60 of eye direction and head position and orientation (pose) [2]. By itself, head
 61 pose provides an estimate that represents a coarse approximation of gaze direction
 62 that can be used in situations in which the eyes are invisible (e.g., when observing
 63 a distant person, or when sunglasses occlude the eyes) [3]. When the eyes are not
 64 occluded, the head pose is an extra marker that can be used to estimate the direction
 65 of the gaze. The gaze direction estimation problem, as it is solved by the human brain,
 66 can therefore be subdivided into two fundamental and *sequential* subproblems: *head*
 67 *pose estimation* and *eye gaze estimation*.

68 The consequences of such a solution are twofold: partial information can be used
 69 to already arrive to an estimate; however, this happens at the expense of biasing. As



Fig. 3 Wollaston illusion: although the eyes are the same in both images, the perceived gaze direction is dictated by the orientation of the head. (Adapted from [2, 3])

70 an illustration of this drawback, in Fig. 3 is shown [2] that the interpretation of the
 71 gaze for an observer is deviated in the direction of the head. In any case, the error
 72 propagated by erroneously estimating one of the features is greatly compensated by
 73 the fact that the human brain is able to yield an estimate *even when only presented*
 74 *with partial or incomplete information*. Moreover, visual features used to detect a
 75 face or an eye do not need to be the same for both cases, so they can be detected
 76 independently, which makes the problem more tractable.

77 Consequently, the following paragraphs will present a summarized survey of solutions
 78 for each subproblem.

79 In the survey by [3], solutions for head pose estimation are divided into eight
 80 categories: seven represent pure methods, while the remaining are hybrid methods,
 81 i.e., combinations of the other methods. The article ends by presenting a quantitative
 82 comparison of the performance of these methods.

83 As mentioned in this survey, most of the computer vision based head pose cal-
 84 culation algorithms have diverged greatly from the results of psychophysical exper-
 85 iments as to how the brain tackles this problem. In fact, the former are concentrated
 86 on *appearance-based* methods, while the latter takes into account how the human
 87 perceives the pose of the head based on *geometrical cues* [3].

88 Geometrical approaches, as shown in Fig. 4, attempt to detect head features as
 89 accurately as possible in order to compute the pose of the head. An example of a
 90 geometrical approach for head pose estimation is presented in [4], where mono-
 91 ocular images are used as input information. The proposed algorithm makes mini-
 92 mal assumptions, compared with other methods, about the facial features structure.
 93 Knowing the positions of the nose, eyes, and mouth, the facial normal direction can
 94 be obtained from one of the next two methods [4], also used in our work:

- 95 1. Using two relations: the nose tip and the line between the far corners of the mouth
 96 ($R_1 = \frac{l_m}{l_f}$); the line between one eye with the correspondent far corners of the
 97 mouth and the distance given by the nose tip; and the line connecting one eye
 98 with the far corners of the mouth ($R_2 = \frac{l_e}{l_f}$);
- 99 2. Using the line between the eye extremities and the far mouth corners.

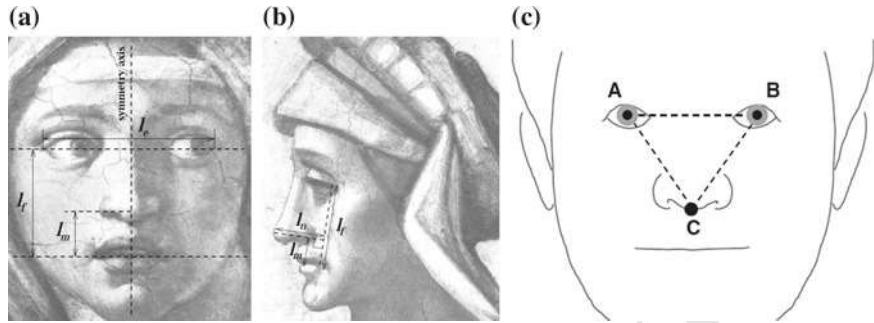


Fig. 4 Geometrical relations between facial features (Adapted from [4]). 3D gaze orientations can be computed using the distances between detected facial features, such as the eyes, nose and mouth

100 The derivation of the roll, pitch, and yaw for a human head is presented in [5]. The
 101 assumption from this article is that the four points that describe the eye are collinear.
 102 The position is obtained using the line through the four eye points and the nose tip.
 103 The main difficulties with this method are related to the pitch direction estimation,
 104 which uses an anthropometric face analysis [5]. The yaw and the pitch are obtained
 105 from eye corners and the intrinsic camera parameters (focal length).

106 The method proposed by [6] uses the model of the face and the eye, deduced from
 107 anthropometric features in order to determine the head orientation. This method uses
 108 only three points (e.g., eye centers and the middle point between the nostrils) to per-
 109 form the desired task. Their model uses the following assumption: $d(A, C) = d(B, C)$;
 110 $d(A, B) = kd \cdot d(A, C)$; $d(A, B) = 6, 5$ cm, where A and B are the central points of
 111 each eye and C is the middle point between the nostrils.

112 Another solution for head pose estimation is introduced in [7]. The main idea
 113 here is to consider an isosceles triangle, with corners in both eyes and in the center
 114 of the mouth. The direction of the head is computed if we assume that one side of
 115 the triangle lies on the image plane, such that applying a trigonometric function we
 116 can estimate the angle between the triangle plane and the image plane [7].

117 Finally, an alternative method for head estimation is supposed to use multiple
 118 cameras [8] with accurate calibration information available. Skin color segmentation
 119 is performed on each camera, and then data fusion is performed, resulting in a 3D
 120 model of the head. The orientation of the head is estimated based on a particle filter.

121 2.2 Facial Features Extraction

122 Feature detection represents a subtopic within the head pose estimation problem. An
 123 accurate estimate for the eye, nose, or the mouth represents an intermediate stage, in
 124 which essential information used by the geometrical approach for head pose estima-
 125 tion is computed. Methods for gaze estimation, presented in the following section,

126 include eye feature detection. Detection of other important facial features, such as
127 the mouth and the nose, is discussed next.

128 Mouth recognition is dealt with methods such as the ones suggested in [9, 10]. A
129 common approach for detecting the mouth is by pre-segmenting the color red on a
130 specific patch of the image. Both methods use a ROI (Region of Interest) extracted
131 after head segmentation, in which the mouth is approximately segmented, after a
132 color space conversion is performed (such as RGB to HSI (*Hue, Saturation, Inten-*
133 *sity*) [9], or RGB to *Lab* [10]). On the other hand, nose detection algorithms use
134 Boosting classifiers, commonly trained with Haar-like features [11], or the 3D infor-
135 mation of the face, as in [12].

136 As suggested in [13], most of the methods used for eyes detection and segmen-
137 tation can be divided into shape-based, appearance-based and hybrid methods. The
138 shape-based technique uses the detection of the iris, the pupil, or the eyelids to locate
139 the eye. Particular features, such as the pupil (dark/bright pupil region) or cornea
140 reflections are used in appearance-based approaches, while the hybrid method tries
141 to combine the advantages of both methods.

142 The shape-based algorithm proposed in [14], built on the isophote curvature con-
143 cept, i.e., the curve that connects points of the same intensity, is able to deliver accu-
144 rate eye localization from a web camera. The main advantage of using this concept
145 is that the shape of the isophotes is invariant to rotation or to linear illumination
146 changes. The eye location can be determined using a combination of Haar features,
147 dual orientation Gabor filters and eye templates, as described in [15].

148 Unsupervised learning algorithms, such as the *Independent Component Analysis*
149 (ICA), are used in [16] for eyes extraction, based on the fact that the eye is a sta-
150 ble facial feature. The two stages technique determines first a rough eye ROI using
151 ICA and the gray-level image intensity variance, and second, the eye center point is
152 computed from image intensity data.

153 Finally, an alternative method which uses two visual sensors is proposed in [17]:
154 a wide-angle camera for face detection and rough eyes estimation and an active pan–
155 tilt–zoom camera to focus on the rough detected ROIs. The method considers the
156 face as a 3D terrain surface and the eye areas as "pits" and "hillsides" regions. The
157 eyes 2D positions are chosen using a (GMM). A similar dual stereo camera system
158 is also proposed in [18], where a wide-angle camera detects the face and an active
159 narrow *Field of View* (FoV) system tracks the eyes at high resolution.

160 As mentioned above, most methods tackle the problem of gaze direction esti-
161 mation using either head pose or eyes direction estimation. However, papers such
162 as [14, 19, 20] present hybrid approaches that combine head pose and eye direction
163 estimation for obtaining the subject's gaze direction.

164 In [14], a hybrid solution for eye detection and tracking, combining the detec-
165 tion results with a *Cylindrical Head Model* (CHM) for head direction estimation,
166 is presented. In [19], the gaze's direction is computed in two stages, after a camera
167 calibration process: first the eyes orientation vector is determined with respect to the
168 head's coordinate system and, second, the final gaze direction estimate is given by
169 a fusion between the determined eyes and head's poses. Both approaches have lim-

170 itations in estimating the gaze's orientation when either the eyes or the poses of the
171 head are imprecise.

172 The technique from [20] describes a human gaze direction algorithm from a com-
173 bination of *Active Appearance Models* (AAM) and a CHM. Although the approach
174 seems to perform well in off-line experiments, real-time scenarios are not presented.
175 One other notable facial features extractor is the Flandmark system [21], which,
176 despite its real-time capabilities and ability to detect and track facial features from
177 frontal faces, fails to recognize features when the pose of the head has a slight offset
178 from the frontal view.

179 3 Controlling a Machine Vision System

180 In a robotics application, the purpose of the machine vision system is to perceive the
181 environment through a camera module.

182 An image processing chain is usually composed of low (e.g., image enhancement,
183 segmentation) and high (e.g., object recognition) level image processing methods.
184 In order for the high level operations to perform properly, the low level ones have
185 to deliver reliable information. In other words, object recognition methods require
186 reliable input coming from previous operations [22].

187 In order to improve the image processing chain, we propose to control the low
188 level vision operation through a feedback loop derived from the higher level compo-
189 nents. In [23, 24], the inclusion of feedback structures within vision algorithms for
190 improving the overall robustness of the chain is suggested.

191 The core idea of the feedback control system for adapting the low level vision
192 operations is presented in Fig. 5, where the control signal u , or *actuator variable*, is
193 a parameter which controls the processing method, whereas the *controlled variable*
194 y is a measure of image processing quality.

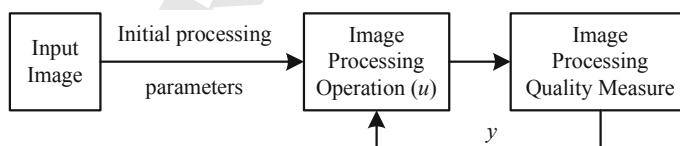


Fig. 5 Feedback adaptation of a computer vision algorithm. The image processing quality measure y is used as a feedback control variable for adapting the parameters of the vision algorithms using the actuator u

195 4 Image Processing Chain

196 The gaze following image processing chain, depicted in Fig. 6, contains four main
 197 steps. We assume that the input is an 8-bit gray-scale image $I = J^{V \times W}$, of width V
 198 and height W , containing a face viewed either from a frontal or profile direction,
 199 where $J = \{0, \dots, 255\}$. (v, w) represents the 2D coordinates of a specific pixel. The
 200 face region is obtained from a face detector.

201 First, a set of facial features ROI hypotheses $\mathbf{H} \in \{h_{le}, h_{re}, h_n, h_m\}$, consisting
 202 of possible instances of the left h_{le} and right h_{re} eyes, nose h_n and mouth h_m , are
 203 extracted using a local features estimator which determines the probability measure
 204 $p(\mathbf{H}|I)$ of finding one of the searched local facial region. The number of computed
 205 ROI hypotheses is governed by a probability threshold T_h , which rejects hypotheses
 206 with a low $p(\mathbf{H}|I)$ confidence measure. The choice of the T_h threshold is not a trivial
 207 task when considering time critical systems, such as the gaze estimator, which, for
 208 a successful HRI, has to deliver in real-time the 3D gaze orientation of the human
 209 subject. The lower T_h is, the higher the computation time. On the other hand, an
 210 increased value for T_h would reject possible “true positive” facial regions, thus leading
 211 to a failure in gaze estimation. As explained in the following, in order to obtain a
 212 robust value for the hypotheses selection threshold, we have chosen to adapt T_h with
 213 respect to the confidences provided by the subsequent estimators from Fig. 6, which
 214 take as input the facial regions hypotheses. The output probabilities coming from
 215 these estimation techniques, that is, the spatial estimator and the GMM for point-
 216 wise feature extraction, are used in a feedback manner within the extremum seeking
 217 control paradigm.

218 Once the hypotheses vector \mathbf{H} has been built, the facial features are combined into
 219 the spatial hypotheses $\mathbf{g} = g_0, g_1, \dots, g_n$, thus forming different facial region combi-
 220 nations. Since one of the main objectives of the presented algorithm is to identify
 221 facial points of frontal, as well as profile faces, a spatial vector s_i is composed either
 222 from four, or three, facial ROIs:

Editor Proof

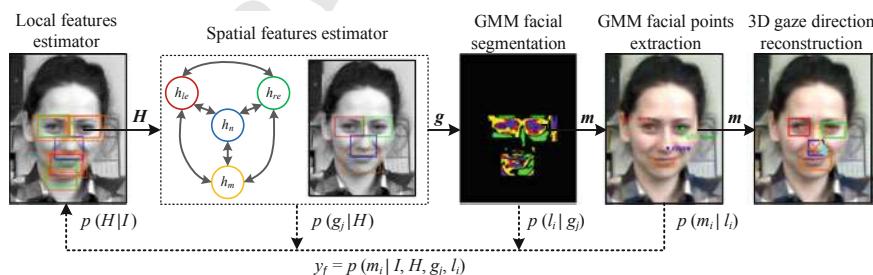


Fig. 6 Block diagram of the proposed gaze following system for facial feature extraction and 3D gaze orientation reconstruction. Each processing block within the cascade provides a measure of feature extraction quality, fused within the controlled variable y_f (see Eq. 2)

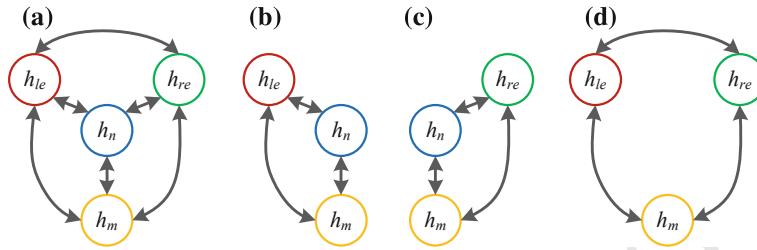


Fig. 7 Different spatial combinations of features used for training the four classifiers. **a** All four facial features. **b, c, d** Cases where only three features are visible in the sample image

$$g_i = \{h_0, h_1, h_2, h_3\} \cap \{h_0, h_1, h_2\}, \quad (1)$$

where $h_i \in \{h_{le}, h_{re}, h_n, h_m\}$.

The extraction of the best spatial features combination can be seen as a graph search problem $g_j = f : G(\mathbf{g}, \mathbf{E}) \rightarrow \mathfrak{R}$, where \mathbf{E} are the edges of the graph connecting the hypotheses in \mathbf{g} . The considered features combinations are illustrated in Fig. 7. Each combination has a specific spatial probability value $p(g_j|\mathbf{H})$ given by a spatial estimator trained using the spatial distances between the facial features from a training database.

Once the spatial distributions of the probable locations of the facial features ROIs are available, their pointwise location m_i is determined using a GMM segmentation method. Its goal is to extract the most probable facial pointwise locations m_i given the GMM pixel likelihood values $p(l_i|g_j)$. The most relevant point features for computing the 3D gaze of a person are the centers of the eyes, tip of the nose, and corners of the mouth.

The described data analysis methods are used to evaluate a feature space composed of the local and spatial features.

Having in mind the facial feature points extraction algorithm described above, it can be stated that the confidence value y_f of the processing chain in Fig. 6 is a probability confidence measure obtained from the estimators cascade:

$$y_f = p(m_i|I, \mathbf{H}, g_j, l_i). \quad (2)$$

Since the whole described processing chain is governed by a set of parameters, such as the threshold T_h for selecting the vector \mathbf{s} , we have chosen to adapt it using an extremum seeking control mechanism and the feedback variable y_f , derived from the output of the gaze following structure illustrated in Fig. 6. The final 3D gaze orientation vector $\vec{\varphi}(m_i)$, representing the roll, pitch, and yaw of the human subject, is determined using the algorithm proposed in the work of Gee and Cipolla [4].

249 5 Performance Evaluation

250 5.1 Experimental Setup

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

253 The system has been evaluated on the *Labeled Faces in the Wild* (LFW) data-
 254 base [25]. LFW consists of 13,233 images, each having a size of $250 \times 250\text{px}$. In
 255 addition to the LFW database, the system has been evaluated on an Adept Pioneer®
 256 3-DX mobile robot equipped with an RGB-D sensor delivering $640\text{px} \times 480\text{px}$ size
 257 color and depth images. The goal of the scenarios is to track the facial features of the
 258 human subject in the HRI context. The error between the real and estimated facial
 259 feature's locations was computed offline.

260 For evaluation purposes, two metrics have been used:

- 261 • the mean normalized deviation between the ground truth and the estimated posi-
 262 tions of the facial features:

$$263 \quad d(\mathbf{m}, \hat{\mathbf{m}}) = \tau(\mathbf{m}) \frac{1}{k} \sum_{i=0}^{k-1} \|m_i - \hat{m}_i\|, \quad (3)$$

264 where k is the number of facial features, \mathbf{m} and $\hat{\mathbf{m}}$ are the manually and estimated
 265 annotated positions of the eyes, nose and mouth, respectively, and $\tau(\mathbf{m})$ is a nor-
 266 malization constant:

$$267 \quad \tau(\mathbf{m}) = \frac{1}{\|(m_{le} + m_{re}) - m_m\|}. \quad (4)$$

- 268 • the maximal normalized deviation:

$$269 \quad d^{\max}(\mathbf{m}, \hat{\mathbf{m}}) = \tau(\mathbf{m}) \max_{j=0, \dots, k-1} \|m_j - \hat{m}_j\|. \quad (5)$$

Editor Proof

270 5.2 Competing Detectors

271 The proposed gaze following system has been tested against three open source detec-
 272 tors.

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

277 The boosting cascades, one for each facial feature, have been trained using a

278 few hundred samples for each eye, nose, and mouth. The searching has been
 279 performed several times at different scales, with Haar-like features used as inputs
 280 to the basic classifiers within the cascade. From the available ROI hypotheses,
 281 the one having the maximum confidence value has been selected as the final
 282 facial feature.

- 283 (2) *Active Shape Models*: An *Active Shape Model* (ASM) calculates a set of feature
 284 points along the facial features contours of the eyes, nose, mouth, eyebrows, or
 285 chin. An ASM is initially trained using a set of manually marked contour points.
 286 The open source AsmLib, based on OpenCV, has been used as candidate detec-
 287 tor. The ASM is trained using manually marked face contours. The trained ASM
 288 model determines variations in the training dataset using *Principal Component
 289 Analysis* (PCA), which enables the algorithm to estimate if the contour is a face.
 290 (3) *Flandmark*: *Flandmark* [21] is a deformable part model detector of facial fea-
 291 tures, where the detection of the point features is treated as an instance of struc-

Editor Proof

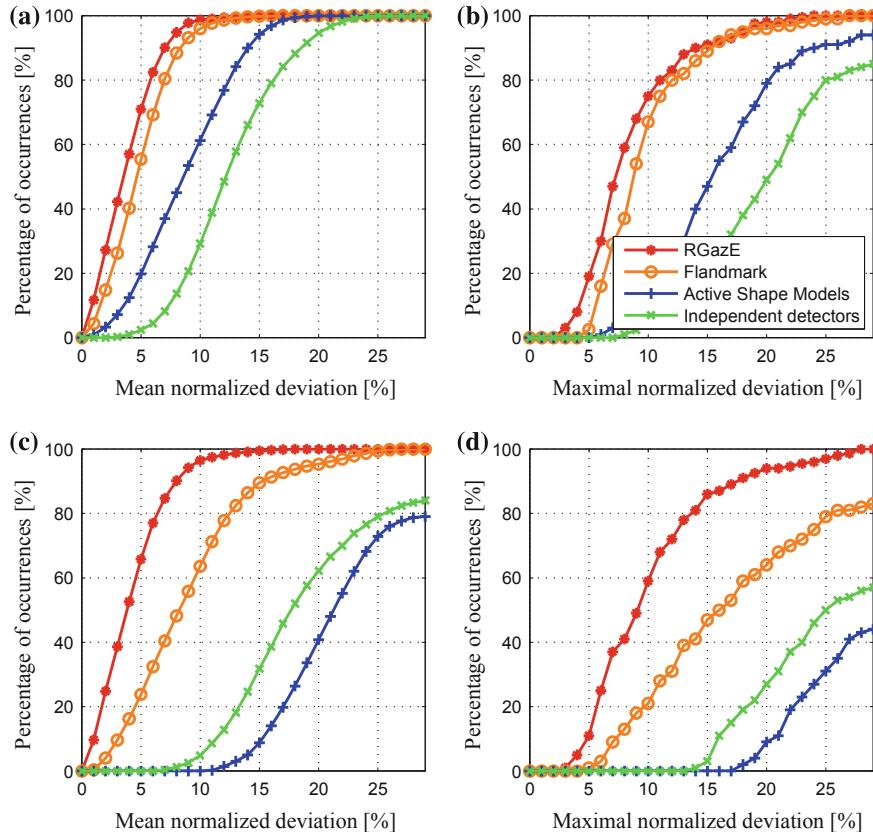


Fig. 8 Cumulative histograms for the mean and the maximal normalized deviation shown for all competing detectors applied on video sequences with frontal (**a**, **b**) and profile (**c**, **d**) faces

292 tured output classification. The algorithm is based on a *Structured Output Support
293 Vector Machine* (SO-SVM) classifier for the supervised learning of the
294 parameters for facial points detection from examples.

295 In comparison to our gaze following system, which uses a segmentation step for
296 determining the pointwise location of the facial features, flandmark considers the
297 centers of the detected ROIs as the point location of the eyes, nose, and mouth.

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

302 The cumulative histograms of the mean and maximal normalized deviation are
303 shown in Fig. 8 for frontal and profile faces. In all cases, the proposed estimator
304 delivered an accuracy value superior to the ones given by the competing detectors.
305 If the accuracy difference between our algorithm and Flandmark is relatively low for
306 the case of frontal faces, it actually increases when the person's face is imaged from
307 a profile view.

308 An interesting observation can be made when comparing the independent detectors
309 with the ASM one. Although the ASM outperforms independent facial feature
310 extraction on frontal faces, it does not perform well when the human subjects are
311 viewed from the lateral. This is due to the training nature of the ASM, where the
312 input training data is made of points spread on the whole frontal area (e.g., eyes,
313 eyebrows, nose, chin, cheeks, etc.).

Editor Proof

314 6 Conclusion

315 In this paper, a robust facial features detector for 3D gaze orientation estimation has
316 been proposed. The solution is able to return a reliable gaze estimate, even if only a
317 partial set of facial features is visible. The paper brings together algorithms for facial
318 feature detection, machine learning, and control theory. During the experiments, we
319 investigated the system's response and compare the results to ground truth values. As
320 shown in the experimental results section, the method performed well with respect
321 to various testing scenarios. As future work, the authors consider the possibility of
322 extending the framework for the simultaneous gaze estimation of multiple interlocu-
323 tors and the adaptation of algorithm with respect to the robot's egomotion.

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

326 References

- 327 1. Scassellati, B.: Theory of mind for a humanoid robot. *Auton. Robots* **12**(1999), 13–24 (2002)
- 328 2. Langton, S.R.H., Honeyman, H., Tessler, E.: The influence of head contour and nose angle on
- 329 the perception of eye-gaze direction. *Atten. Percept. Psychophys.* **66**(5), 752–771 (2004)
- 330 3. Chotorian, E., Trivedi, M.: Head pose estimation in computer vision: a survey. *IEEE Trans.*
- 331 *Pattern Anal. Mach. Intell.* **31**(4), 607–629 (2009)
- 332 4. Gee, A., Cipolla, R.: Determining the gaze of faces in images. *Image Vis. Comput.* **12**(10),
- 333 639–647 (1994)
- 334 5. Horprasert, T., Yacoob, Y., Davis, L.: Computing 3-d head orientation from a monocular image
- 335 sequence. In: *Proceedings of the Second International Conference on Automatic Face and Gesture*
- 336 *Recognition*, pp. 242–247, Oct 1996
- 337 6. Kaminski, J., Knaan, D., Shavit, A.: Single image face orientation and gaze detection. *Mach.*
- 338 *Vis. Appl.* **21**(3), 85–98 (2009)
- 339 7. Nikolaidis, A., Pitas, I.: Facial feature extraction and pose determination. *Pattern Recogn.*
- 340 **33**(11), 1783–1791 (2000)
- 341 8. Canton-Ferrer, C., Casas, J., Pardas, M.: Head orientation estimation using particle filtering in
- 342 multiview scenarios. In: *Multimodal Technologies for Perception of Humans*, vol. 4625, pp.
- 343 317–327. Springer, Berlin (2008)
- 344 9. Pantic, M., Tomc, M., Rothkrantz, L.: A hybrid approach to mouth features detection. In:
- 345 *2001 IEEE International Conference on Systems, Man, and Cybernetics*, vol. 2, pp. 1188–1193
- 346 (2001)
- 347 10. Skodras, E., Fakotakis, N.: An unconstrained method for lip detection in color images. In: *2011*
- 348 *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1013–1016
- 349 (2011)
- 350 11. Gonzalez-Ortega, D., Diaz-Pernas, F., Martinez-Zarzuela, M., Anton-Rodriguez, M., Diez-
- 351 Higuera, J., Boto-Giralda, D.: Real-time nose detection and tracking based on adaboost and
- 352 optical flow algorithms. In: *Intelligent Data Engineering and Automated Learning*, vol. 5788,
- 353 pp. 142–150. Springer, Berlin (2009)
- 354 12. Werghi, N., Boukadia, H., Meguebli, Y., Bhaskar, H.: Nose detection and face extraction from
- 355 3d raw facial surface based on mesh quality assessment. In: *36th Annual Conference on IEEE*
- 356 *Industrial Electronics Society*, pp. 1161–1166 (2010)
- 357 13. Hansen, D., Ji, Q.: In the eye of the beholder: a survey of models for eyes and gaze. *IEEE*
- 358 *Trans. Pattern Anal. Mach. Intell.* **32**(3), 78–500 (2010)
- 359 14. Valenti, R., Sebe, N., Gevers, T.: Combining head pose and eye location information for gaze
- 360 estimation. *IEEE Trans. Image Process.* (2011)
- 361 15. Ke, L., Kang, J.: Eye location method based on haar features. In: *2010 3rd International*
- 362 *Congress on Image and Signal Processing*, vol. 2, pp. 925–929 (2010)
- 363 16. Hassaballah, M., Kanazawa, T., Ido, S.: Efficient eye detection method based on grey intensity
- 364 variance and independent components analysis. *Comput. Vis. IET* **4**(4), 261–271 (2010)
- 365 17. Reale, M., Canavan, S., Yin, L., Hu, K., Hung, T.: A multi-gesture interaction system using a
- 366 3-d iris disk model for gaze estimation and an active appearance model for 3-d hand pointing.
- 367 *IEEE Trans. Multimedia* **13**(3), 474–486 (2011)
- 368 18. Beymer, D., Flickner, M.: Eye gaze tracking using an active stereo head. In: *2003 IEEE Com-*
- 369 *puter Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 451–458
- 370 (2003)
- 371 19. Ronssse, R., White, O., Lefevre, P.: Computation of gaze orientation under unrestrained head
- 372 movements. *J. Neurosci. Methods* **159**, 158–169 (2007)
- 373 20. Sung, J., Kanade, T., Kim, D.: Pose robust face tracking by combining active appearance mod-
- 374 els and cylinder head models. *Int. J. Comput. Vis.* **80**, 260–274 (2008)
- 375 21. Uřičář, M., Franc, V., Hlaváč, V.: Detector of facial landmarks learned by the structured output
- 376 SVM. In: Csurka, G., Braz, J. (eds.) *VISAPP '12: Proceedings of the 7th International Confer-*
- 377 *ence on Computer Vision Theory and Applications*, vol. 1, pp. 547–556. SciTePress—Science
- 378 and Technology Publications, Portugal, Feb 2012

- 379 22. Hotz, L., Neumann, B., Terzic, K.: High-level expectations for low-level image processing. In:
380 KI 2008: Advances in Artificial Intelligence. Springer, Berlin (2008)
- 381 23. Ristic, D.: Feedback structures in image processing. Ph.D. dissertation, Bremen University,
382 Institute of Automation, Bremen, Germany, Apr 2007
- 383 24. Grigorescu, S.M.: Robust machine vision for service robotics. Ph.D. dissertation, Bremen Uni-
384 versity, Institute of Automation, Bremen, Germany, June 2010
- 385 25. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a data-
386 base for studying face recognition in unconstrained environments, University of Massachusetts,
387 Amherst. Technical Report 07-49, Oct 2007

Editor Proof

