FROG

FUN ROBOTIC
OUTDOOR GUIDE

Deliverable: D4.2

Demonstrator of human conversational signals analyzer

Consortium

UNIVERSITEIT VAN AMSTERDAM (UvA)
IDMIND - ENGENHARIA DE SISTEMAS LDA (IDM)
UNIVERSIDAD PABLO DE OLAVIDE (UPO)
IMPERIAL COLLEGE OF SCIENCE, TECHNOLOGY AND MEDICINE (ICL)
UNIVERSITY OF TWENTE (UT)

Grant agreement no. **288235**
Funding scheme **STREP**

SEVENTH FRAMEWORK
PROGRAMME

DOCUMENT INFORMATION

Project

Project acronym:	FROG
Project full title:	Fun Robotic Outdoor Guide
Grant agreement no.:	288235
Funding scheme:	STREP
Project start date:	1 October 2011
Project duration:	36 Months
Call topic:	ICT-2011.2.1 Cognitive Systems and Robotics (a), (d)
Project web-site:	www.frogrobot.eu

Document

Deliverable number:	D4.2
Deliverable title:	Demonstrator of human conversational signals analyzer
Due date of deliverable:	M33 - 31 July 2014
Actual submission date:	M36 – 14 September 2014
Editors:	ICL
Authors:	J. Shen, I. Marras, M. Pantic
Reviewers:	Dariu Gavrila UvA
Participating beneficiaries:	1, 5
Work Package no.:	4
Work Package title:	Socially Adaptive Human Robot Interaction
Work Package leader:	UT
Work Package participants:	1,2,5,6
Estimated person-months for deliverable:	15
Dissemination level:	Public
Nature:	Demonstrator
Version:	2
Draft/Final:	Final
No of pages (including cover):	42
Keywords:	Face verification, face alignment, head nod/shake detection, social attitudes, agreeing/disagreeing

Chapter 1

Introduction

To operate efficiently and in a socially-acceptable manner, a robot in outdoor environments like the FROG robot needs to analyze people's behavior in its environment. Since visitors are seldom alone, it is also essential for the robot to be able to distinguish between individual visitors within the same group. A guide robot needs to be aware of the context in which it interacts with people. Specifically, the robot needs to know which visitor it is interacting with, what location the visitor is interested in, whether he / she is engaged and what information or content the visitor wishes to see. In order to detect the visitor's interest level toward the content provided by the robot, we identify relevant human behaviors including facial expressions and certain conversational cues.

In particular, this involves the detection / recognition of human visitors and the analysis of user behavior from head gestures as well as facial expression. These results enable the robot to interpret the visitor's behavior in terms of his or her level of interest, emotion-related states like engagement, and social signals like agreement and disagreement. Face recognition is also performed to allow the robot to know when a new visitor comes to start interacting with it. Thus the robot would be able to adapt its tour guide strategy based on the new visitor's implicit affective feedback (e.g. attention and interest). The challenges lie in the reliable detection of human behaviors in outdoor environments with changing light conditions and frequent noise and occlusions. Integrating such vision data will allow the identification of human activities as well as user implicit feedback from affective signals. Identifying users' engagement and interests will significantly enhance the robot's effectiveness and acceptance since it allows the robot to exhibit "social awareness" and appropriately respond to relevant events. This is an important step towards a feasible application of robots in outdoor guide scenarios.

The task in this deliverable aims further :

- to develop a method for human face recognition in the wild. As one of the most important biometric techniques, face recognition has a clear advantage of being natural and passive over other biometric techniques that require user cooperation (i.e. fingerprint, iris). The image of the human face can have large intra-subject variations (changes in the same individual), thus making it difficult to develop an accurate recognition system in the wild. The system is supposed to be able to identify / recognize an uncooperative face in uncontrolled environments without the notice of the subject. Face image can be captured from a distance without touching the person being identified, and the identification does not require interaction with the person. The non-intrusive nature of face recognition leads to many challenging issues. In particular, a successful "in the wild" facial recognition method needs:

  - to handle face pose variations,
  - to handle illumination changes,
  - to handle partial occlusion of the face caused by accessories such as lenses, sunglasses and scarves, facial hair (mustache, beard, long hair, etc.), and changes in facial features due to aging,
  - to solve the one sample per person problem in face recognition,
  - to perform fully automatic face recognition.

  So far, many face recognition algorithms have tried to deal with some of the above problems in:

  - well controlled environments,
  - uncontrolled environments.

  Face recognition in well controlled environments is relatively an easier task. thus many approaches proposed in the past performed well under such conditions. However, face recognition remains an unsolved problem under realistic conditions for real applications. As a result of the aforementioned challenges, the problem of accurate face recognition becomes much harder in uncontrolled environments. A very limited number of methods tried to deal with these issues. The method used in this task deals with all these challenges and performs automatic face recognition in uncontrolled (outdoor) environments.

- This task also aims: to develop a multi-cue visual method for detection of social attitudes like agreeing and disagreeing. Detection is based on the state of the art in cognitive sciences and based on morphological and temporal correlations between relevant visual cues, including facial gestures like frowns and smiles as well as head gestures like tilts and nods. To this aim, a method that deals with noise, clutter and occlusions present in outdoor spaces is used in order to attain agreement and disagreement detection based on automatically detected (rather than manual annotated) behavioral cues. For this purpose, facial landmarks as well as face pose were used. The method for facial landmark detection as well as for face pose estimation was developed for FROG and it was presented in deliverable 3.1 (FROG tracker).

These two main steps are presented in detail in the next chapters.

Chapter 2

Robust and Efficient Parametric Automatic Face Recognition

The task of face recognition in the wild is performed by applying an image registration technique using two images of a person without prior information about this person. More specifically, the core of the proposed algorithm is an image alignment procedure using a piecewise affine motion model. Thus, we propose a new cost function for gradient ascend face recognition: the maximization of the correlation of image gradient orientations. The use of this correlation coefficient has been motivated by the recent success of FFT-based gradient correlation methods for the robust estimation of translational displacements [48, 21]. More specifically, we use a correlation coefficient which takes the form of the sum of cosines of gradient orientation differences. The use of gradient orientation differences is the key to the robustness of the proposed scheme. As it is was shown in [48, 49], local orientation mismatches caused by outliers can be well-described by a uniform distribution which, under a number of mild assumptions, is canceled out by applying the cosine kernel. Thus, image regions corrupted by outliers result in approximately zero correlation and therefore do not bias the estimation of the transformation parameters significantly. To maximize the gradient correlation coefficient, we formulate and solve a continuous optimization problem.

2.1 Gradient-based correlation coefficient

Assume that we are given the image-based representations of two objects $\mathbf{I}_i \in \Re^{m_1 \times m_2}, i = 1, 2$. We define the complex representation which combines the magnitude and the orientation of image gradients as $\mathbf{G}_i = \mathbf{G}_{i,x} + j\mathbf{G}_{i,y}$, where $j = \sqrt{(-1)}$, $\mathbf{G}_{i,x} = \mathbf{F}_x \star \mathbf{I}_i$, $\mathbf{G}_{i,y} = \mathbf{F}_y \star \mathbf{I}_i$ and $\mathbf{F}_x, \mathbf{F}_y$ are filters used to approximate the ideal differentiation operator along the image horizontal and vertical direction respectively. We also denote by $\mathcal{P}$ the set of indices corresponding to the image support and by $\mathbf{g}_i = \mathbf{g}_{i,x} + j\mathbf{g}_{i,y}$ the $N-$dimensional vectors obtained by writing $\mathbf{G}_i$ in lexicographic ordering, where N is the cardinality of $\mathcal{P}$. The gradient correlation coefficient is defined as

$$s \triangleq \Re\{\mathbf{g}_1^H \mathbf{g}_2\}, \tag{2.1}$$

where $\Re\{.\}$ denotes the real part of a complex number and H denotes the conjugate transpose [48]. Using $\mathbf{r}_i(k) \triangleq \sqrt{\mathbf{g}_{i,x}^2(k) + \mathbf{g}_{i,y}^2(k)}$ and $\phi_i(k) \triangleq \arctan \frac{\mathbf{g}_{i,y}(k)}{\mathbf{g}_{i,x}(k)}$, we have

$$s \triangleq \sum_{k \in \mathcal{P}} \mathbf{r}_1(k)\mathbf{r}_2(k) \cos[\Delta\boldsymbol{\phi}(k)], \tag{2.2}$$

where $\Delta\boldsymbol{\phi} \triangleq \boldsymbol{\phi}_1 - \boldsymbol{\phi}_2$.

The magnitudes $\mathbf{r}_i$ in (2.2) suppress the contribution of areas of constant intensity level which do not provide useful features for object alignment. Note, however, that the use of gradient magnitude does not necessarily result in robust algorithms. For example, the authors in [13] have shown that the gradient magnitude varies drastically with the change in the direction of the light source.

The key to the robustness of our scheme is the correlation of gradient orientations which takes the form of the sum of cosines of gradient orientation differences [48, 21]. To show this [48, 49], assume that there exists a subset $\mathcal{P}_o \subset \mathcal{P}$ corresponding to the set of pixels corrupted by outliers, while $\mathcal{P}_1$ denotes the image support that is outlier-free ($\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$). By using the normalized gradients $\tilde{\mathbf{g}}_i = \tilde{\mathbf{g}}_{i,x} + j\tilde{\mathbf{g}}_{i,y}$, where $\tilde{\mathbf{g}}_{i,x}(k) = \mathbf{g}_{i,x}(k)/|\mathbf{g}_i(k)|$ and $\tilde{\mathbf{g}}_{i,y}(k) = \mathbf{g}_{i,y}(k)/|\mathbf{g}_i(k)|$, so that $\mathbf{r}_i(k) = 1 \; \forall k$, the value of this gradient correlation coefficient in $\mathcal{P}_o$ is

$$q_o \triangleq \sum_{k \in \mathcal{P}_o} \cos[\Delta\boldsymbol{\phi}(k)]. \tag{2.3}$$

To compute the value of q_o, we note that in $\mathcal{P}_o$ the images are *visually dissimilar/unrelated*, so that locally do not match. It is therefore not unreasonable to assume that for any spatial location k, the difference in gradient orientation $\Delta\boldsymbol{\phi}(k)$ can take any value in the range $[0, 2\pi)$ with equal probability. Thus, we can assume that $\Delta\boldsymbol{\phi}$ is a realization of a stationary random process $u(t)$ which $\forall t$ follows a uniform distribution $U(0, 2\pi)$. Given this, it is not difficult to show that, under some rather mild assumptions, it holds

$$q_o = \sum_{k \in \mathcal{P}_o} \cos[\Delta\boldsymbol{\phi}(k)] \simeq 0. \tag{2.4}$$

Note that (a) in contrary to [40], no assumption about the structure of outliers is made and (b) no actual knowledge of $\mathcal{P}$ is required. Based on (2.4), we can re-write (2.2) as follows

$$\begin{aligned} s &= \sum_{k \in \mathcal{P}_o} \cos[\Delta\boldsymbol{\phi}(k)] + \sum_{k \in \mathcal{P}_1} \cos[\Delta\boldsymbol{\phi}(k)] \\ &= \sum_{k \in \mathcal{P}_o} \epsilon \cdot \cos[\Delta\boldsymbol{\phi}(k)] + \sum_{k \in \mathcal{P}_1} 1 \cdot \cos[\Delta\boldsymbol{\phi}(k)] \\ &\approx \mathbf{g}_1^T \mathcal{Q}_{ideal} \mathbf{g}_2, \end{aligned} \tag{2.5}$$

where $\epsilon \to 0$ and $\mathcal{Q}_{ideal}$ is the "ideal" weighting matrix defined above. Note that $\mathcal{Q}_{ideal}$ in (2.5) is never calculated explicitly. We can write (2.5) only because outliers are approximately "canceled out" when the above kernel is used to measure image similarity.

This assumption has been shown to be valid using the Kolmogorov-Smirnoff test for more than 70.000 pairs of *visually unrelated images* in [49]. As an example, in Fig. 2.1 (a)-(b), we assume that the scarf is visually unrelated to the face. $\mathcal{P}_o$ here corresponds to the part of the face occluded by the scarf defined by the red rectangle. Fig. 2.1 (c) plots the distribution of $\Delta\phi$ in $\mathcal{P}_o$, while Fig. 2.1 (d) shows the histogram of uniformly distributed samples obtained with Matlab's rand function. As in [49], to verify that $\Delta\phi$ is uniformly distributed, we used the Kolmogorov-Smirnov test [39] to test the null hypothesis $H_0 : \forall k \in \mathcal{P}_o,\ \Delta\phi(k) \sim U[0, 2\pi)$. For a significance level of 0.01, the null hypothesis was accepted with p-value equal to 0.254. Similarly, for the samples obtained with Matlab's `rand` function, the null hypothesis was accepted with $p = 0.48$.

Overall, unlike standard correlation (i.e. the inner product) of pixel intensities where the contribution of outliers can be arbitrarily large, the effect of outliers is approximately canceled out in $\mathcal{P}_o$. Corrupted regions result in approximately zero correlation and thus do not bias the estimation of the transformation parameters.

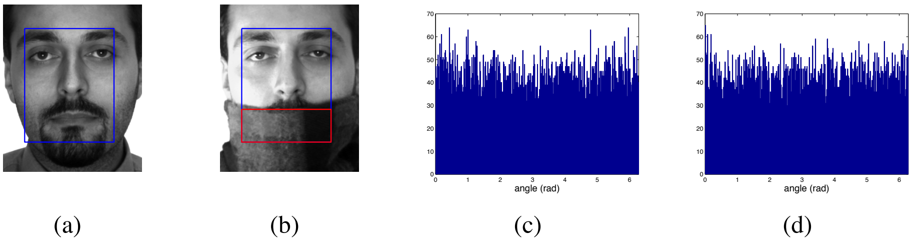


Figure 2.1: (a)-(b) A pair of faces from the AR database. The region of interest is defined by the blue rectangle. The corrupted region $\mathcal{P}_o$ is defined by the red rectangle. (c) The distribution of $\Delta\phi$ in $\mathcal{P}_o$. (d) The distribution of samples (uniformly distributed) obtained with Matlab's rand function.

2.2 Gradient Orientation in Face Analysis

The use of gradient orientation as useful features for face analysis is by no means proposed for the first time in this work. Examples of previous work can be found in [30, 14, 13]. However, most prior work proposes gradient orientations as features for achieving insensitivity in non-uniform illumination variations. On the contrary, what is highlighted in [48, 49] as well as in this work is why gradient orientations can be used for outlier-robust (for example occlusion-robust) face analysis.

Regarding face alignment, perhaps what is somewhat related to the our scheme is the Active Appearance Model proposed in [14]. We underline two important differences between our algorithm and the method of [14]. First, as [14] *does* employ the gradient magnitude (even for normalization) for feature extraction, it is inevitably less robust to outliers. Second, no attempt to exploit the relation between image gradients and pixel intensities is made. More specifically, the gradient-based features in [14] are treated just as pixel intensities which are then used for *regression-based* object alignment. On the contrary, we make full use of the relation between image gradients and pixel intensities to formulate and solve a *continuous optimization* problem. This results in a dramatic performance improvement as Section 2.4 illustrates.

2.3 Robust and efficient object recognition

Our method deals with the most difficult face recognition problem when only one sample is known per person. This problem is very common in FROG data since it is very possible only a few images for some person to be available, while only a few of them depict its face in frontal position without face occlusions or bad face illumination thus these images can be used as a face template during the face recognition scenario. In this case, the creation of subspace models is impossible. Having two images of a person, one for training and one for testing, the correction of misalignment errors between them must be performed by the cost function. In the next sub-sections this cost function will be presented in detail.

Parametric object alignment methods assume that $\mathbf{I}_1$ and $\mathbf{I}_2$ are related by a parametric transformation, i.e.

$$\mathbf{I}_1(\mathbf{x}_k) = \mathbf{I}_2(\mathbf{W}(\mathbf{x}_k;\mathbf{p})), \forall k \in \mathcal{P}, \tag{2.6}$$

where $\mathbf{W}(\mathbf{x}_k;\mathbf{p})$ is the parametric transformation with respect to the image coordinates $\mathbf{x}_k = [\mathbf{x}_1(k), \mathbf{x}_2(k)]^T$ and $\mathbf{p} = [\mathbf{p}(1), \ldots, \mathbf{p}(n)]^T$ is the vector of the unknown parameters. This parametric transformation is based on the motion model used. There are many motion models that can be used. In our case, we utilized two motion models, an affine and a piecewise affine motion model. The unknown parameters $\mathbf{p}$ in the first case are defined by the affine transformation itself. In the second case, motivated by AAMs, a shape model is used. This shape model is typically learned by annotating N fiducial points on the object (e.g. a face) of training image $\mathbf{I}_i$. These points are said to define the shape of each object. Next, Procrustes analysis is applied to remove similarity transformations from the original shapes. Finally, PCA is applied on the similarity-free shapes $s_i = [x_1, y_1, x_2, y_2, \cdots x_N, y_N]$. The resulting model $\{\mathbf{\Phi}_{S,0}, \mathbf{\Phi}_S \in \Re^{2N \times p}\}$ can be used to represent a test shape $\mathbf{s}_y$ as

$$\widehat{\mathbf{s}}_y = \mathbf{\Phi}_{S,0} + \mathbf{\Phi}_S\mathbf{p}, \quad \mathbf{p} = \mathbf{\Phi}_S^T(\mathbf{s}_y - \mathbf{\Phi}_{S,0}). \tag{2.7}$$

The eigenvectors of $\mathbf{\Phi}_S$ represent pose, expression and identity variation. For each of the images $\mathbf{I}_1$ and $\mathbf{I}_2$ the N facial landmarks are detected by using an automatic procedure for facial landmark detection. Then, the alignment procedure is driven by these automatic detected facial landmarks using piecewise affine warp. The image $\mathbf{I}_2(\mathbf{W}(\mathbf{x}_k;\mathbf{p}))$ is computed by backwards warping the input image $\mathbf{I}_2$ with the warp $\mathbf{W}(\mathbf{x}_k;\mathbf{p})$, i.e. for each pixel $\mathbf{x}$ in the base mesh $\mathbf{\Phi}_{S,0}$ we compute $\mathbf{W}(\mathbf{x};\mathbf{p})$ and sample (bilinearly interpolate) the image $\mathbf{I}_2$ at that location.

Next, $\mathbf{p}$ is estimated by minimizing an objective function which is typically the ℓ_2 norm of the difference $\mathbf{E} = \mathbf{I}_1 - \mathbf{I}_2$. The minimization is performed in an iterative fashion after making a first or second order Taylor approximation to either $\mathbf{I}_1$ or $\mathbf{I}_2$. It is obvious that above image difference becomes an image alignment problem.

2.3.1 The quantity maximized

In this section, we introduce the maximization of the correlation of image gradient orientations as a new cost function for robust gradient descent face alignment. In particular, to estimate $\mathbf{p}$, we wish to maximize

$$q = \sum_{k \in \mathcal{P}} \cos[\Delta \boldsymbol{\phi}(k)]. \tag{2.8}$$

By using the normalized gradients $\tilde{\mathbf{g}}_i$, simple calculations show that (2.8) is equivalent to

$$q = \sum_{k \in \mathcal{P}} \tilde{\mathbf{g}}_{1,x}(k)\tilde{\mathbf{g}}_{2,x}(k) + \tilde{\mathbf{g}}_{1,y}(k)\tilde{\mathbf{g}}_{2,y}(k). \tag{2.9}$$

Note, however, that a first order Taylor expansion of $\tilde{\mathbf{g}}_1$ or $\tilde{\mathbf{g}}_2$ with respect to $\Delta\mathbf{p}$ yields a linear function of $\Delta\mathbf{p}$ which is maximized as $\Delta\mathbf{p} \to \infty$. To alleviate this problem without resorting to the second order Taylor expansion as in [50], we follow an approach similar to [19]. To proceed, we note that as $||\tilde{\mathbf{g}}_2(k)||_2 = 1, \forall k \in \mathcal{P}$, the cost function is exactly equal to

$$q = \frac{\sum_{k \in \mathcal{P}} \tilde{\mathbf{g}}_{1,x}(k)\tilde{\mathbf{g}}_{2,x}(k) + \tilde{\mathbf{g}}_{1,y}(k)\tilde{\mathbf{g}}_{2,y}(k)}{\sqrt{\sum_{k \in \mathcal{P}} \tilde{\mathbf{g}}_{2,x}^2(k) + \tilde{\mathbf{g}}_{2,y}^2(k)}}, \tag{2.10}$$

but if we linearize $\tilde{\mathbf{g}}_2$ in the above expression, the denominator will not be equal to 1 and q will become a non-linear function of $\Delta\mathbf{p}$. Finally, using vector notation, our cost function becomes

$$q = \frac{\tilde{\mathbf{g}}_{1,x}^T\tilde{\mathbf{g}}_{2,x} + \tilde{\mathbf{g}}_{1,y}^T\tilde{\mathbf{g}}_{2,y}}{\sqrt{\tilde{\mathbf{g}}_{2,x}^T\tilde{\mathbf{g}}_{2,x} + \tilde{\mathbf{g}}_{2,y}^T\tilde{\mathbf{g}}_{2,y}}}. \tag{2.11}$$

To maximize q with respect to $\mathbf{p}$, we first make the dependence of $\tilde{\mathbf{g}}_2(k)$ on $\mathbf{p}$ explicit by writing $\tilde{\mathbf{g}}_2[\mathbf{p}](k)$. Then, we maximize iteratively by assuming that the current estimate of $\mathbf{p}$ is known and by looking for an increment $\Delta\mathbf{p}$ which maximizes our objective function in (2.11) with respect to $\Delta\mathbf{p}$.

2.3.2 The forward-additive gradient correlation algorithm

In this section, we describe how to maximize our cost function in (2.11) using the forward-additive maximization procedure. In this framework [35, 6], at each iteration, we maximize (2.11) with respect to $\Delta\mathbf{p}$ where $\mathbf{g}_2 \longleftarrow \mathbf{g}_2[\mathbf{p} + \Delta\mathbf{p}]$. Once we obtain $\Delta\mathbf{p}$, we update the parameter vector in an additive fashion $\mathbf{p} \longleftarrow \mathbf{p} + \Delta\mathbf{p}$ and use this new value of $\mathbf{p}$ to obtain the updated warped image $\mathbf{I}_2(\mathbf{W}(\mathbf{x}; \mathbf{p}))$.

We start by noting that $\mathrm{g}_2[\mathbf{p}](k)$ is the complex gradient of $\mathbf{I}_2(\mathbf{W}(\mathbf{x};\mathbf{p}))$ with respect to the original coordinate system evaluated at $\mathbf{x}=\mathbf{x}_k$. This gradient is different from the gradient of $\mathbf{I}_2$ calculated at the first iteration and then evaluated at $\mathbf{W}(\mathbf{x}_k;\mathbf{p})$, which, for convenience, we will denote by $\mathbf{h}_2[\mathbf{p}](k)$. That is, $\mathbf{h}_2[\mathbf{p}]=\mathbf{h}_{2,x}[\mathbf{p}]+j\mathbf{h}_{2,y}[\mathbf{p}]$ is obtained by writing $\mathbf{G}_{2,x}(\mathbf{W}(\mathbf{x};\mathbf{p}))+j\mathbf{G}_{2,y}(\mathbf{W}(\mathbf{x};\mathbf{p}))$ in lexicographic ordering, where $\mathbf{G}_2=\mathbf{G}_{2,x}+j\mathbf{G}_{2,y}$ is assumed to be computed at the first iteration. In a similar fashion, we denote by $\mathbf{h}_{2,xx}[\mathbf{p}]$, $\mathbf{h}_{2,yy}[\mathbf{p}]$ and $\mathbf{h}_{2,xy}[\mathbf{p}]$, the vectors obtained by writing in lexicographic ordering the second partial derivatives of I_2, $\mathbf{G}_{2,xx}$, $\mathbf{G}_{2,yy}$ and $\mathbf{G}_{2,xy}$, computed at the first iteration and, then, evaluated at $\mathbf{W}(\mathbf{x};\mathbf{p})$. Let us also write $\mathbf{W}(\mathbf{x};\mathbf{p})=[\mathbf{w}_1(\mathbf{x};\mathbf{p}),\mathbf{w}_2(\mathbf{x};\mathbf{p})]^T$, so that the matrix derivative $\frac{\partial\mathbf{W}}{\partial\mathbf{a}}$ with respect to a vector $\mathbf{a}=[\mathbf{a}(1),\ldots,\mathbf{a}(m)]^T$ depends on the motion model used. The derivative $\frac{\partial\mathbf{W}}{\partial\mathbf{a}}$ is given as follows:

- in case of affine motion model the derivative is given by

$$\frac{\partial\mathbf{W}}{\partial\mathbf{a}}=\begin{bmatrix}\frac{\partial\mathbf{w}_1}{\partial\mathbf{a}(1)} & \cdots & \frac{\partial\mathbf{w}_1}{\partial\mathbf{a}(m)}\\ \frac{\partial\mathbf{w}_2}{\partial\mathbf{a}(1)} & \cdots & \frac{\partial\mathbf{w}_2}{\partial\mathbf{a}(m)}\end{bmatrix}. \tag{2.12}$$

- in case of a piecewise affine motion model, the calculation and implementation of the derivative can be found in [37].

By definition we have

$$\begin{aligned}\mathrm{g}_2[\mathbf{p}](k) &\triangleq [\mathrm{g}_{2,x}[\mathbf{p}](k)\ \ \mathrm{g}_{2,y}[\mathbf{p}](k)]\\ &\triangleq \left.\frac{\partial\mathbf{I}_2(\mathbf{W}(\mathbf{x};\mathbf{p}))}{\partial\mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_k}\\ &= \nabla_{\mathbf{W}}\mathbf{I}_2[\mathbf{p}](k)\left.\frac{\partial\mathbf{W}}{\partial\mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_k},\end{aligned} \tag{2.13}$$

where $\nabla_{\mathbf{W}}\mathbf{I}_2[\mathbf{p}](k)\triangleq[\mathbf{h}_{2,x}[\mathbf{p}](k)\ \ \mathbf{h}_{2,y}[\mathbf{p}](k)]$. By applying the chain rule and noticing that $\nabla_{\mathbf{W}}\frac{\partial\mathbf{W}}{\partial\mathbf{x}}=0$, we also have

$$\begin{aligned}\begin{bmatrix}\frac{\partial\mathrm{g}_{2,x}[\mathbf{p}](k)}{\partial\mathbf{p}}\\ \frac{\partial\mathrm{g}_{2,y}[\mathbf{p}](k)}{\partial\mathbf{p}}\end{bmatrix} &= \left(\left.\frac{\partial\mathbf{W}}{\partial\mathbf{x}}\right|_{\mathbf{x}=\mathbf{x}_k}\right)^T\\ &\times\begin{bmatrix}\mathbf{h}_{2,xx}[\mathbf{p}](k) & \mathbf{h}_{2,xy}[\mathbf{p}](k)\\ \mathbf{h}_{2,yx}[\mathbf{p}](k) & \mathbf{h}_{2,yy}[\mathbf{p}](k)\end{bmatrix}\frac{\partial\mathbf{W}}{\partial\mathbf{p}}.\end{aligned} \tag{2.14}$$

We assume that the current estimate of $\mathbf{p}$ is known. The key point to make derivations tractable is to recall that $\tilde{\mathrm{g}}_{2,x}[\mathbf{p}](k)\equiv\cos\phi_2[\mathbf{p}](k)$ and $\tilde{\mathrm{g}}_{2,y}[\mathbf{p}](k)\equiv\sin\phi_2[\mathbf{p}](k)$ where

$$\phi_2[\mathbf{p}](k)=\arctan\frac{\mathrm{g}_{2,y}[\mathbf{p}](k)}{\mathrm{g}_{2,x}[\mathbf{p}](k)}. \tag{2.15}$$

By performing a first order Taylor expansion on $\tilde{\mathbf{g}}_{2,x}[\mathbf{p}+\Delta\mathbf{p}](k)$, we get

$$\tilde{\mathbf{g}}_{2,x}[\mathbf{p}+\Delta\mathbf{p}](k) \approx \cos\boldsymbol{\phi}_2[\mathbf{p}](k) + \frac{\partial\cos\boldsymbol{\phi}_2[\mathbf{p}](k)}{\partial\mathbf{p}}\Delta\mathbf{p}. \tag{2.16}$$

By repeatedly applying the chain rule, we get

$$\frac{\partial\cos\boldsymbol{\phi}_2[\mathbf{p}](k)}{\partial\mathbf{p}} = -\sin\boldsymbol{\phi}_2[\mathbf{p}](k)\mathbf{j}[\mathbf{p}](k), \tag{2.17}$$

where $\mathbf{j}[\mathbf{p}](k)$ is a $1\times n$ vector given by

$$\mathbf{j}[\mathbf{p}](k) = \frac{\cos\boldsymbol{\phi}_2[\mathbf{p}](k)\frac{\partial\mathbf{g}_{2,y}[\mathbf{p}](k)}{\partial\mathbf{p}} - \sin\boldsymbol{\phi}_2[\mathbf{p}](k)\frac{\partial\mathbf{g}_{2,x}[\mathbf{p}](k)}{\partial\mathbf{p}}}{\sqrt{\mathbf{g}_{2,x}^2[\mathbf{p}](k)+\mathbf{g}_{2,y}^2[\mathbf{p}](k)}}. \tag{2.18}$$

Using vector notation, we can write

$$\tilde{\mathbf{g}}_{2,x}[\mathbf{p}+\Delta\mathbf{p}] \approx \cos\boldsymbol{\phi}_2[\mathbf{p}] - \mathbf{S}_\phi[\mathbf{p}]\odot\mathbf{J}[\mathbf{p}]\Delta\mathbf{p}, \tag{2.19}$$

where $\mathbf{S}_\phi[\mathbf{p}]$ is the $N\times n$ matrix whose $k-$th row has n elements all equal to $\sin\boldsymbol{\phi}_2[\mathbf{p}](k)$, $\mathbf{J}[\mathbf{p}]$ is the $N\times n$ Jacobian matrix whose $k-$th row has n elements corresponding to $\mathbf{j}[\mathbf{p}](k)$ and $\odot$ denotes the Hadamard product. Very similarly, we can derive

$$\tilde{\mathbf{g}}_{2,y}[\mathbf{p}+\Delta\mathbf{p}] \approx \sin\boldsymbol{\phi}_2[\mathbf{p}] + \mathbf{C}_\phi[\mathbf{p}]\odot\mathbf{J}[\mathbf{p}]\Delta\mathbf{p}, \tag{2.20}$$

where $\mathbf{C}_\phi[\mathbf{p}]$ is the $N\times n$ matrix whose $k-$th row has n elements all equal to $\cos\boldsymbol{\phi}_2[\mathbf{p}](k)$.

Let us denote by $\mathbf{S}_{\Delta\phi}[\mathbf{p}]$ the $N\times 1$ vector whose $k-$th element is equal to $\sin(\boldsymbol{\phi}_1(k)-\boldsymbol{\phi}_2[\mathbf{p}](k))$. Then, by plugging (2.19) and (2.20) into (2.11), and after some calculations, our cost function becomes

$$q(\Delta\mathbf{p}) = \frac{q_\mathbf{p} + \mathbf{S}_{\Delta\phi}^T\mathbf{J}\Delta\mathbf{p}}{\sqrt{N+\Delta\mathbf{p}^T\mathbf{J}^T\mathbf{J}\Delta\mathbf{p}}}, \tag{2.21}$$

where $q_\mathbf{p} = \cos\boldsymbol{\phi}_1^T\cos\boldsymbol{\phi}_2 + \sin\boldsymbol{\phi}_1^T\sin\boldsymbol{\phi}_2$ is the correlation of gradient orientations between $\mathbf{I}_1$ and $\mathbf{I}_2(\mathbf{W}(\mathbf{x};\mathbf{p}))$, and we have dropped the dependence of the quantities on $\mathbf{p}$ for notational simplicity. Finally, the maximization of (2.21) with respect to $\Delta\mathbf{p}$ can be obtained by applying the results of [19]. In particular, the maximum value is attained for

$$\Delta\mathbf{p} = \lambda(\mathbf{J}^T\mathbf{J})^{-1}\mathbf{J}^T\mathbf{S}_{\Delta\phi}, \tag{2.22}$$

where $\lambda = \frac{1}{\tilde{q}}$ and $\tilde{q} = q_\mathbf{p}/N$ denotes the normalized correlation (such that $|\tilde{q}|\leq 1$) Thus, λ has a very intuitive interpretation. As $\tilde{q}$ is small (large) in the first (last) iterations, a large (small) λ is used as a weight in (2.22).

2.3.3 The inverse-compositional gradient correlation algorithm

In this section, we show how to maximize our cost function in (2.11) using the inverse-compositional maximization procedure. In this framework [5, 6], a change of variables is made to switch the roles of $\mathbf{I}_1$ and $\mathbf{I}_2$ and the updated warp is obtained in a compositional (rather than additive) fashion. Thus, our cost function becomes

$$q = \frac{(\tilde{\mathbf{g}}_{2,x}[\mathbf{p}])^T(\tilde{\mathbf{g}}_{1,x}[\Delta\mathbf{p}]) + (\tilde{\mathbf{g}}_{2,y}[\mathbf{p}])^T(\tilde{\mathbf{g}}_{1,y}[\Delta\mathbf{p}])}{\sqrt{(\tilde{\mathbf{g}}_{1,x}[\Delta\mathbf{p}])^T(\tilde{\mathbf{g}}_{1,x}[\Delta\mathbf{p}]) + (\tilde{\mathbf{g}}_{1,y}[\Delta\mathbf{p}])^T(\tilde{\mathbf{g}}_{1,y}[\Delta\mathbf{p}])}} \tag{2.23}$$

with respect to $\Delta\mathbf{p}$ and, at each iteration, $\mathbf{I}_2$ is updated using $\mathbf{W}(\mathbf{x};\mathbf{p}) \longleftarrow \mathbf{W}(\mathbf{x};\mathbf{p}) \circ (\mathbf{W}(\mathbf{x};\Delta\mathbf{p}))^{-1}$, where $\circ$ denotes composition.

Similarly to [6], we assume that $\mathbf{W}(\mathbf{x};\mathbf{0}) = \mathbf{x}$. This, in turn, implies $\mathbf{g}_1[\Delta\mathbf{p}] \equiv \mathbf{h}_1[\Delta\mathbf{p}]$ which greatly simplifies the derivations. As before, we perform a Taylor approximation to $\tilde{\mathbf{g}}_{1,x}[\mathbf{p}]$, but this time around zero. This gives

$$\tilde{\mathbf{g}}_{1,x}[\Delta\mathbf{p}] \approx \cos\boldsymbol{\phi}_1[\mathbf{0}] - \mathbf{S}_\phi[\mathbf{0}] \odot \mathbf{J}[\mathbf{0}]\Delta\mathbf{p}, \tag{2.24}$$

where $\mathbf{S}_\phi[\mathbf{0}]$ is the $N\times n$ matrix whose $k-$th row has n elements all equal to $\sin\boldsymbol{\phi}_1[\mathbf{0}](k)$ and $\mathbf{J}[\mathbf{0}]$ is the $N \times n$ Jacobian matrix whose $k-$th row has n elements corresponding to the $1 \times n$ vector

$$\mathbf{j}[\mathbf{0}](k) = \frac{\cos\boldsymbol{\phi}_1[\mathbf{0}](k)\frac{\partial \mathbf{g}_{1,y}[\mathbf{0}](k)}{\partial \mathbf{p}} - \sin\boldsymbol{\phi}_1[\mathbf{0}](k)\frac{\partial \mathbf{g}_{1,x}[\mathbf{0}](k)}{\partial \mathbf{p}}}{\sqrt{\mathbf{g}_{1,x}^2[\mathbf{0}](k) + \mathbf{g}_{1,y}^2[\mathbf{0}](k)}} \tag{2.25}$$

and

$$\begin{bmatrix} \frac{\partial \mathbf{g}_{1,x}[\mathbf{0}](k)}{\partial \mathbf{p}} \\ \frac{\partial \mathbf{g}_{1,y}[\mathbf{0}](k)}{\partial \mathbf{p}} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_{1,xx}[\mathbf{0}](k) & \mathbf{g}_{1,xy}[\mathbf{0}](k) \\ \mathbf{g}_{1,yx}[\mathbf{0}](k) & \mathbf{g}_{1,yy}[\mathbf{0}](k) \end{bmatrix} \left.\frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right|_{\mathbf{p}=\mathbf{0}}.$$

Similarly, for $\tilde{\mathbf{g}}_{1,y}[\Delta\mathbf{p}]$, we get

$$\tilde{\mathbf{g}}_{1,y}[\Delta\mathbf{p}] \approx \sin\boldsymbol{\phi}_1[\mathbf{0}] + \mathbf{C}_\phi[\mathbf{0}] \odot \mathbf{J}[\mathbf{0}]\Delta\mathbf{p}, \tag{2.26}$$

where $\mathbf{C}_\phi[\mathbf{0}]$ is the $N\times n$ matrix whose $k-$th row has n elements all equal to $\cos\boldsymbol{\phi}_1[\mathbf{0}](k)$. Notice that all terms in (2.24) and (2.26) do not depend on $\mathbf{p}$ and, thus, are pre-computed and constant across iterations.

Let us denote by $\mathbf{S}_{\Delta\phi}[\mathbf{p}]$ the $N\times 1$ vector whose $k-$th element is equal to $\sin(\boldsymbol{\phi}_2[\mathbf{p}](k) - \boldsymbol{\phi}_1(k))$. Then, by dropping the dependence of the above quantities on $\mathbf{p}$ and $\mathbf{0}$, our objective function will be again given by (2.21) while the optimum $\Delta\mathbf{p}$ will be given by (2.22).

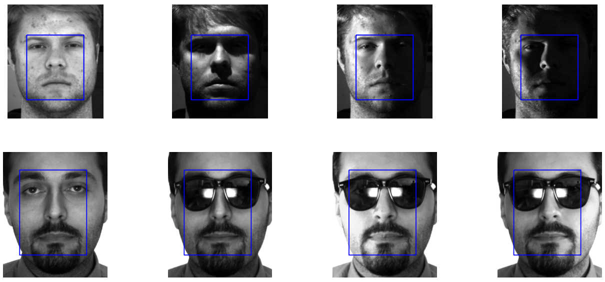


Figure 2.2: Examples of images used in our experiments (prior to the application of an affine transformation). The blue rectangle defines the region of interest.

Methods	Number of image pairs considered	Real image pair	Transformation Affine/ Homography	Illumination	Occlusion	AWGN	Compared with
[6]	4 (Takeo+3)	No	Yes/Yes	No	No	Yes	[6]
[3]	6 (Takeo)	No	Yes/No	No	Yes (synthetic)	No	[6, 3]
[16]	3	Yes	Yes/No	Yes (natural)	No	No	[6]
[19]	1 (Takeo)	No	Yes/No	Yes (synthetic)	No	Yes	[6, 4]
[2]	NA (Multi-Pie [26])	Yes	Yes/No	Yes (natural)	No	No	[6]
[38]	11	No	Yes/No	Yes (synthetic)	No	Yes	[6]
Ours	182 (Takeo + Yale +AR)	Yes	Yes/No	Yes (natural)	Yes (real)	Yes	[6, 3, 19, 2]

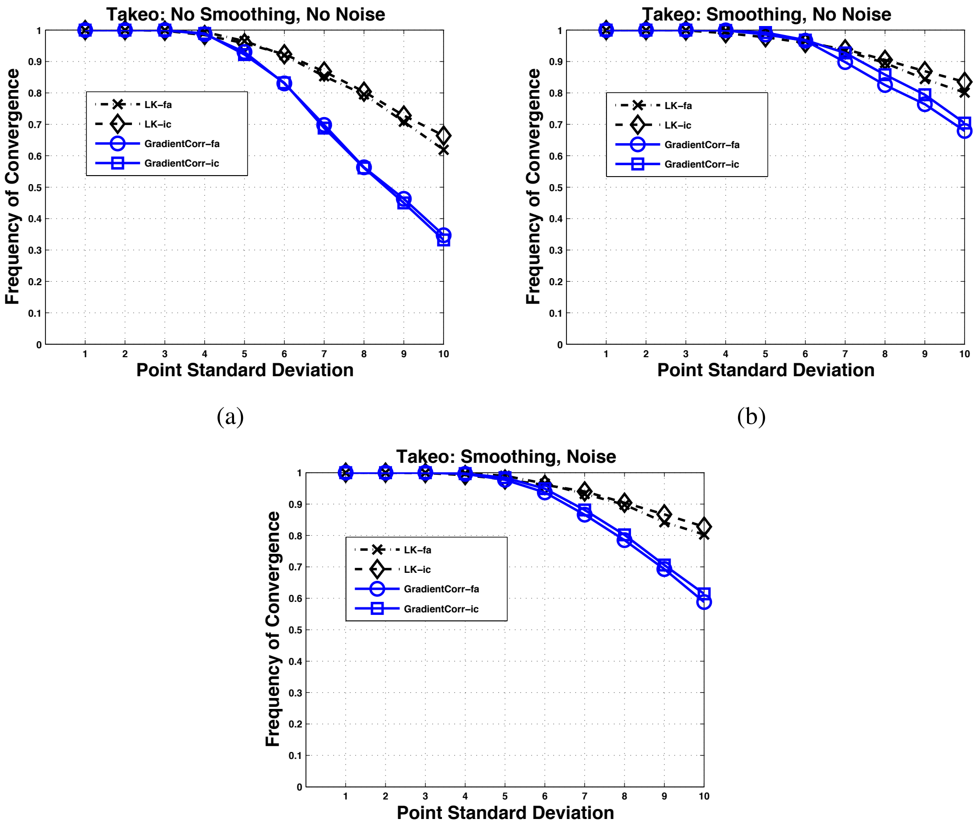
Table 2.1: Comparison between the experimental settings reported in object alignment papers following the evaluation framework of [6].

2.4 Experimental results

Let us define two variations of our framework for face alignment coined as GradientCorr-FA and GradientCorr-IC. We have conducted two different sets of experiments. Since a face alignment step is performed during the face recognition procedure, face alignment experiments are performed separately from the experiments for face recognition. Both method variations were used in the first set of experiments, while, after judged the results of these experiments, GradientCorr-IC were used in the second set of experiments. The difference between these two sets of experiments is the motion model used. More specifically, an affine motion model was used for face alignment experiments, while a piecewise affine motion model was used for the face recognition experiments. These experiments are very important in order to show that our framework does not depend on the motion model used. Below, all these experiments are presented in detail.

2.4.1 Face alignment experiments

We assessed the performance of the image alignment step, used in our face recognition method, using the performance evaluation framework proposed in [6] which has now become the standard evaluation procedure [19, 16, 2, 38]. We present results and comparison with previous work for very challenging alignment cases which have not been previously examined. Table 2.1 presents a comparison between our experiments



(a)

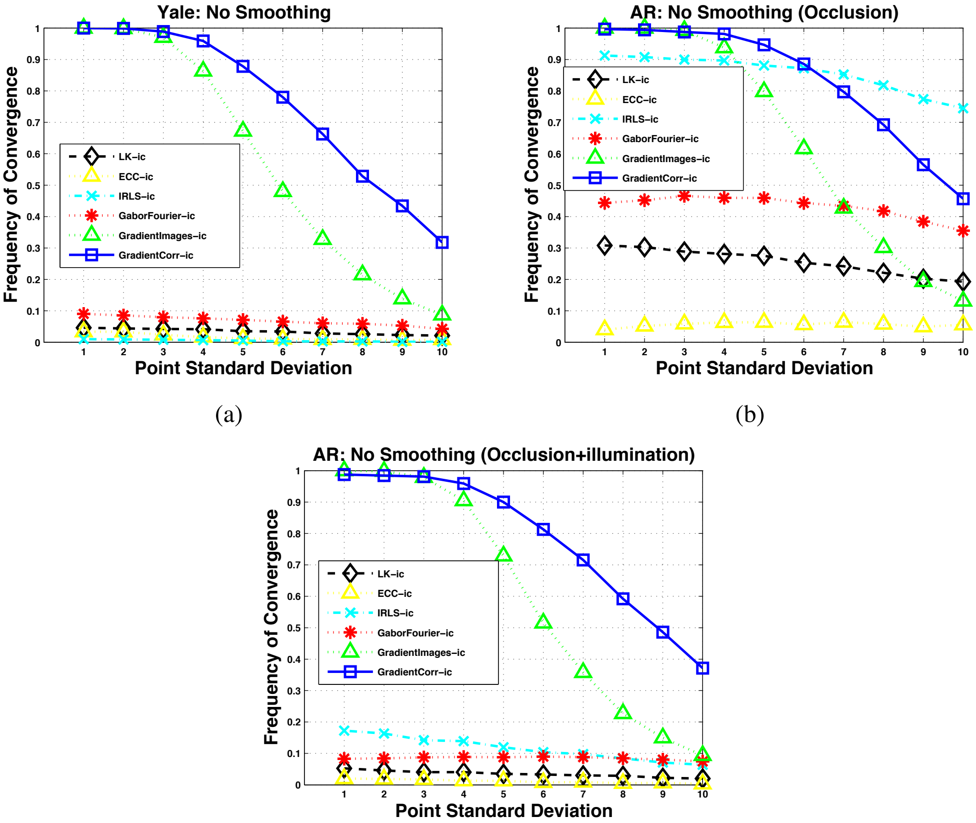
(b)

(c)

Figure 2.3: Frequency of Convergence vs Point Standard Deviation for Takeo image [6]. (a) No Smoothing, No Noise (b) Smoothing, No Noise (c) Smoothing, Noise. LK-fa: black-x. LK-IC: black-◇ . GradientCorr-fa: blue-○. GradientCorr-IC: blue-□.

and the ones reported in object alignment papers which also adopt the evaluation framework of [6]. In addition to the standard "Takeo" experiment, we considered, for the first time (to the best of our knowledge), the problem of face alignment in the presence of real occlusions and non-uniform illumination changes using hundreds of real faces taken from the AR [36] and Yale B [23] databases.

The evaluation in [6] is as follows. We selected a region of interest and three canonical points in this region. We perturbed these points using Gaussian noise of standard deviation σ and computed the initial RMS error between the canonical and perturbed points. Using the affine warp that the original and perturbed points defined, we generated the affine distorted image. Given a warp estimate, we computed the destination of the three canonical points and, then, the final RMS error between the estimated and correct locations. We used the average rate of convergence for a fixed σ and the average frequency of convergence for $\sigma = [1, 10]$ as the performance evaluation measures. An



(a)

(b)

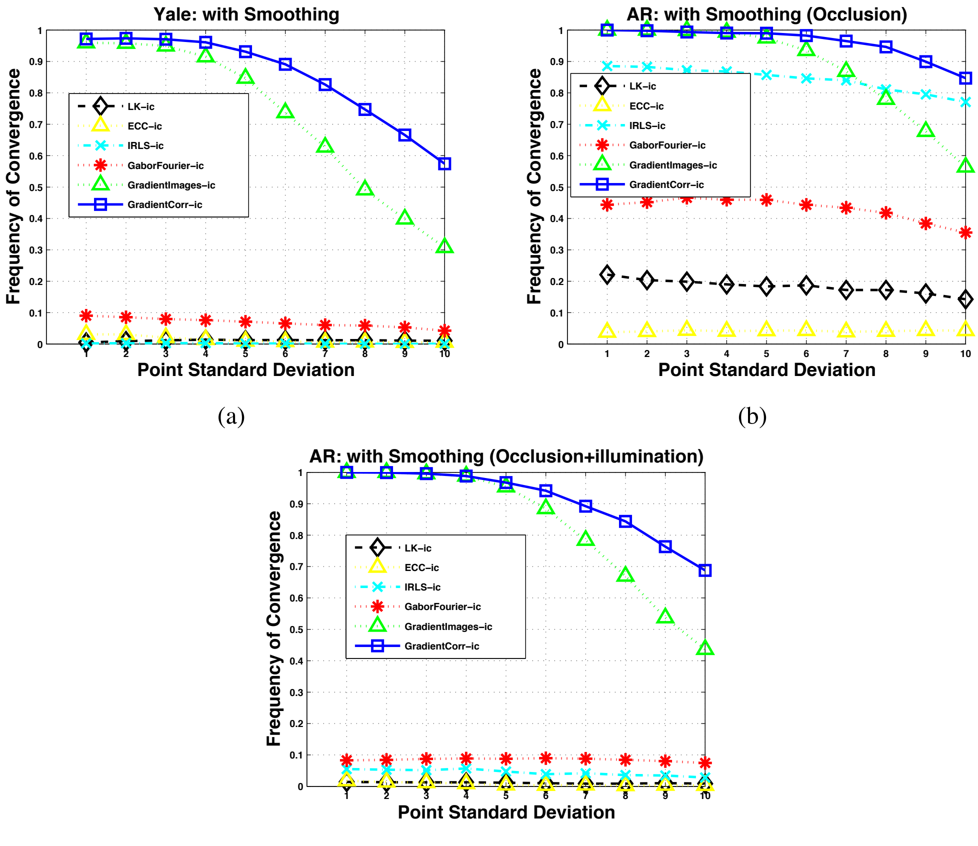
(c)

Figure 2.4: Average Frequency of Convergence vs Point Standard Deviation for Yale and AR databases. **No smoothing was used**. (a) Yale (b) AR-Occlusion (c) AR-Occlusion+illumination. LK-IC: black-◇. ECC-IC: yellow-△. IRLS-IC: cyan-x. GaborFourier-IC: red-*. GradientImages-IC: green-△. GradientCorr-IC: blue-□.

algorithm was considered to have converged if the final RMS point error was less than n_1 pixels after 30 iterations. We obtained these averages using, for each σ, n_2 randomly generated warps.

Experiments using the Takeo image

We started by reproducing to some extend the experimental setting of [6] using the Takeo image. We used $n_1 = 1$ pixel and, for each σ, $n_2 = 1000$ randomly generated warps. We assessed the performance of the forward additive and inverse compositional versions of our algorithm and the LK algorithm. We considered 3 cases. The first case was with no Gaussian smoothing prior to the calculation of image derivatives and no AWGN (Additive White Gaussian Noise). The second case was with smoothing but no AWGN. Finally, the third case was with both smoothing and AWGN of variance equal to 10 added to both the template and the target image. Fig. 2.3 shows the obtained



(a)

(b)

(c)

Figure 2.5: Average Frequency of Convergence vs Point Standard Deviation for Yale and AR databases. **Smoothing was used**. (a) Yale (b) AR-Occlusion (c) AR-Occlusion+illumination. LK-IC: black-◇. ECC-IC: yellow-△. IRLS-IC: cyan-x. GaborFourier-IC: red-*. GradientImages-IC: green-△. GradientCorr-IC: blue-□.

average frequency of convergence.

As Fig. 2.3 (a) shows, for this experiment, the LK algorithms outperform the proposed methods. This is not unreasonable, as the affine distorted image was generated directly from the original image. In this case, there are no outliers, and as our algorithms remove some amount of information (most importantly the gradient magnitude), they inevitably perform worse. As Fig. 2.3 (b) illustrates, Gaussian smoothing improves the performance of all methods by providing a larger region of attraction. The performance gap between the LK and the proposed methods is now significantly smaller. Finally, as Fig. 2.3 (c) shows, if smoothing is used, none of the methods is affected too much by the AWGN even for a large noise variance (In fact, the performance of the LK methods is not affected at all). However, as next section shows, smoothing will not increase the robustness of methods which are not designed to be robust.

Experiments on the Yale and AR databases

In this section, we present our performance evaluation results obtained by using real image pairs (manually aligned), taken from the Yale B [23] and AR databases [36]. Our target was to assess performance in the presence of non-uniform illumination changes and occlusions. We used 100 different face pairs taken from the Yale database as follows. For each of the 10 subjects of the database we selected 1 template and 10 test images corrupted by extreme illumination changes. We also used 81 different face pairs taken from the AR database as follows. We selected 27 out of 31 subjects from the "dbf1" folder (4 subjects were discarded due to significant pose variation). For each subject, we selected 1 template image and 3 test images with sunglasses. Fig. 2.2 shows examples of images used in our experiments.

We used the average frequency of convergence for $\sigma = [1, 10]$ as the performance evaluation measure. We used $n_1 = 3$ pixels and, for each σ, $n_2 = 100$ randomly generated warps. Thus, for each σ, we used a total of 100×100 and 81×100 warps for Yale and AR respectively.

We assessed the performance of the inverse compositional versions of our algorithm (GradientCorr-IC), the LK algorithm (LK-IC) [6], the enhanced correlation (ECC-IC) algorithm [19], the iteratively re-weighted least squares algorithm (IRLS-IC) [3], and the Gabor-Fourier LK algorithm (GaborFourier-IC) recently proposed in [2]. The last two methods as well the mutual-information LK [16] (not considered here) are previously proposed robust methods. The implementations of the LK-IC and IRLS-IC algorithms are kindly provided by the authors. We implemented ECC-IC based on the forward additive implementation of ECC which is also kindly provided by the corresponding authors. Finally, we implemented GaborFourier-IC based on the implementation of LK-IC.

Additionally, based on the discussion in Section 2.2, we propose a new method: we used the orientation-based features of [14] and replaced regression with the inverse compositional algorithm. As gradients are treated exactly the same as intensities, we call this algorithm GradientImages-IC. We included this algorithm in our experiments to illustrate the performance improvement achieved by our scheme which solves a continuous optimization problem based on the relation between gradients and intensities.

With the exception of GaborFourier-IC, for all methods, we considered two cases. The first case was with no Gaussian smoothing while the second one was with smoothing prior to the calculation of the image derivatives. We did not use smoothing for GaborFourier-IC as this is already incorporated in the method.

Figs. 2.4 and 2.5 show the average frequency of convergence for all face pairs and algorithms considered for the cases of "No Smoothing" and "Smoothing" respectively.

Overall, the proposed GradientCorr-IC largely outperformed all other methods resulting in the most robust and stable performance. The performance improvement compared to GradientImages-IC is also more than evident. In particular, for large σ, GradientCorr-IC converged approximately 30-40$\%$ more frequently than GradientImages-IC. As Fig. 2.5 shows, Gaussian smoothing improved the performance of GradientCorr-IC and GradientImages-IC only. IRLS-IC seems to have worked well in the presence of occlusions but failed to converge when illumination changes were present. Surprisingly, Gaussian smoothing reduced the algorithm's performance. Although the results of [2] demonstrate that GaborFourier-IC is much more robust than the original LK-IC algorithm, our results show that this algorithm was also not able to cope with the extreme illumination conditions and occlusions considered in our experiments. Finally, the LK-IC and ECC-IC algorithms are not robust and, not too surprisingly, diverged for almost all face pairs considered.

Computational complexity

A simple inspection of our algorithms shows that the most computationally expensive step is the calculation of $\mathbf{J}^T\mathbf{J}$ in (2.21) which requires $O(n^2N)$ operations. The cost of all other steps is at most $O(nN)$ (since $N \gg n$). In the inverse compositional maximization procedure, $\mathbf{J}^T\mathbf{J}$ and its inverse is pre-computed and, therefore, the complexity per iteration is $O(nN)$. Finally, an un-optimized MATLAB version of our algorithm takes about 0.03-0.04 seconds per iteration while the original inverse compositional algorithm takes about 0.02-0.03 seconds per iteration. We note that an optimized version of the original inverse compositional algorithm, as the core part of Active Appearance Model fitting, has been shown to track faces faster than 200 fps [25].

2.4.2 Face recognition experiments

Face recognition experiments were conducted on three well-known databases: FRGC [41], MULTI-PIE [26] and FERET [42, 43]. Note that face recognition using only one training sample per person is a difficult problem for many methods based on subspace learning proposed in the past. In order to asses the performance of GradientCorr-IC, it is compared to the state-of-the-art method [12] described before.

Both methods require the detection of facial landmarks. To tackle the problem of recognising faces under real conditions when only a single sample per classes is available, an automatic state-of-the-art method for facial landmark detection was used [53]. Thus, the fiducial points in both the training and the testing samples were detected in an automatic way.

The method [12] uses several high dimensional features extracted on facial landmarks (salient points) in multiple scales. In total, 27 landmarks of the inner face were selected as proposed by the authors. As features we used: Histogram of Oriented Gradients (HOG) [15], Scale-invariant feature transform (SIFT) [34], Local Binary Patterns (LBP)

[1] as well as the combination of the three of them. For simplicity reasons let us define us HDF-HOG, HDF-SIFT, HDF-LBP and HDF-Fusion the four different configurations of this method used in our experiments. As mentioned before for the proposed method, the piecewise affine motion motion is driven by N fiducial points on the face. Although the prior shape model is typically learned by manually annotation of these points, during the face recognition under real conditions this is not possible. Furthermore, N=68 fiducial points in total were used, as depicted in Figure 2.6. Figure 2.7 depicts examples of automatic detected facial landmarks using images used in our experiments. As it is depicted in this figure, automatic detection of facial landmarks produces in some cases inaccurate results, especially in cases of facial expressions. This problem is realistic in real face recognition scenarios, thus it has to be treated in a special way by an automatic face recognition method. All the face recognition experiments are described in detail in the next sub-sections.

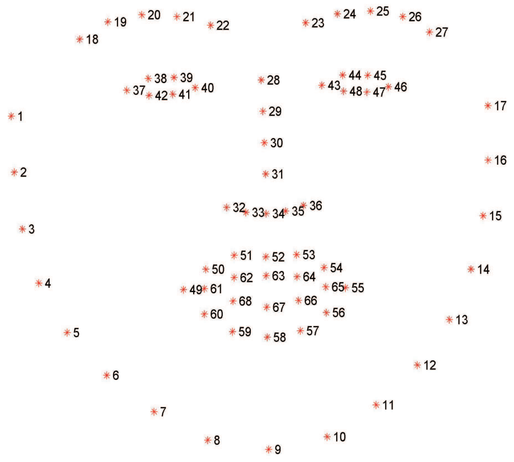


Figure 2.6: The 68 fiducial face points mark-up used to create our shape model.

Experiments on FRGC database

The FRGC v2 database contains 4007 records of 466 persons. The records contain various facial expressions (e.g., happiness, surprise), and subjects are 57% male and 43% female, with the following age distribution: 65% 18-22 years old, 18% 23-27 and 17% 28 years or over. In the FRGC three masks are defined over the square similarity matrix which holds the similarity values between facial records. A mask selects a subset of the records to be used as the gallery set and another subset to be used as the probe set. In the verification scenario, each probe is compared to one gallery set and the result is compared against a threshold, which is one-to-one matching. The results are summarized using Receiver Operating Characteristic (ROC) curves. Each mask is used to perform a different verification experiment, thus producing three different ROC curves, which will be referred to as ROCI, II and III. In ROC I all the data are within

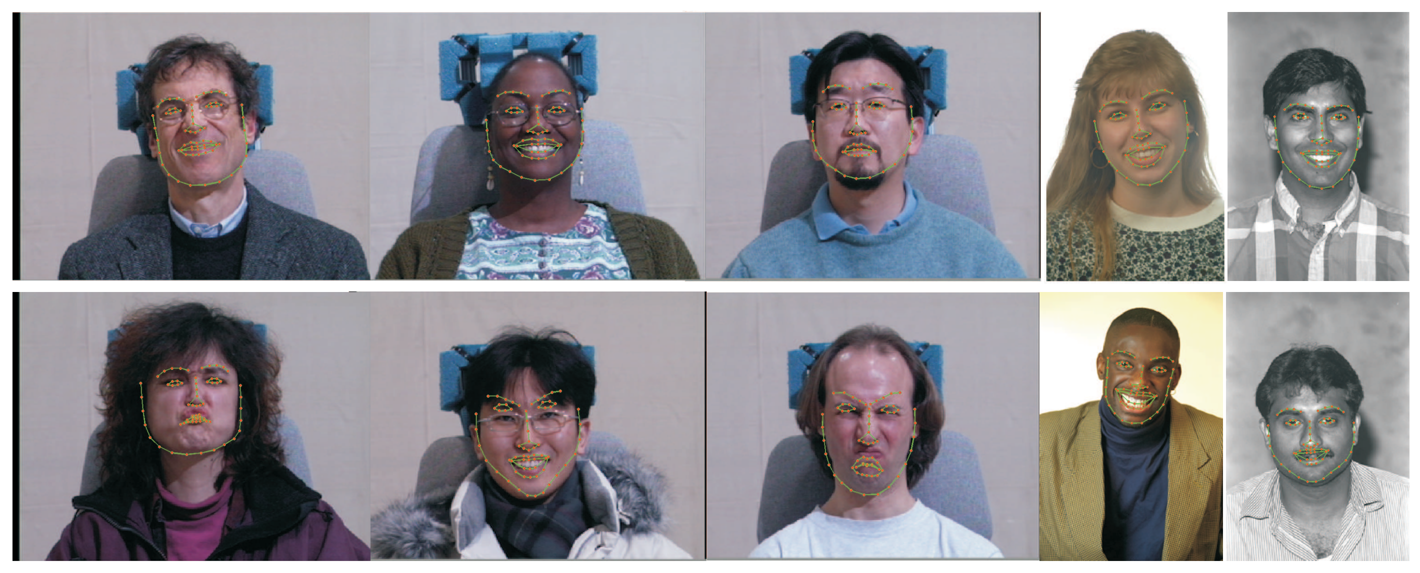


Figure 2.7: Examples of automatic detected facial landmarks using images used in our experiments. In the first row the facial landmarks are well detected. Second row depicts cases where the landmarks are not well detected, thus this problem has to be treated in a special way by an automatic face recognition method.

semesters, in ROC II they are within a year, while in ROC III the samples are between semesters. These experiments are of increasing difficulty.

In our experiment we used the most challenging third verification scenario and measured the verification rate and the false acceptance rate (FAR) and summarize the results on ROC curves. For FAR=0%, GradientCorr-IC achieves verification rate of around 79.50%, while the best configuration of the method in [12] for this database, HDF-HOG, achieves verification rate of around 75.50%. This can be seen in Table 2.2. For FAR=0.1%, GradientCorr-IC achieves verification rate of around 96.50%, while HDF-HOG achieves verification rate of around 94.80%. For the standard protocol test ROC III mask of FRGC v2, the results achieved by the tested methods are summarized in Figure 2.8.

Experiments on FERET database

This collection of images consists of hard recognition cases that have proven difficult for most face recognition algorithms previously tested on the FERET database. The difficulty posed by this data set appears to stem from the fact that the images were taken at different times, at different locations, and under different imaging conditions. We carried out single-sample-per-class face recognition experiments on the FERET database. The evaluation methodology requires that the training must be performed using the FA

Table 2.2: Automatic face verification experiment in FRGC v2 database using the most challenging third verification scenario for FAR=0%.

Methods	Recognition accuracy FAR=0%
HDF-Fusion	73.8411%
HDF-HOG	75.5440%
HDF-SIFT	75.3548%
HDF-LBP	73.8411%
GradientCorr-IC	**79.4702%**

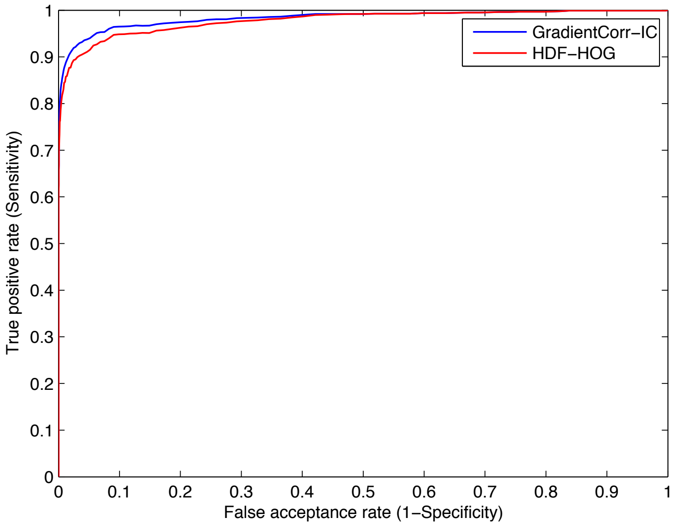


Figure 2.8: ROC curves based on the experiment in FRGC database.

set which contains one frontal view per subject and in total 1196 subjects. No other data set was used for training. The testing sets include the FB, DupI and DupII data sets. Since current techniques achieve almost 100% recognition performance on FB, we used only Dup I and II in our experiments. DupI and DupII probe sets contain 727 and 234 test images, respectively, captured significantly later than FA. These data sets are very challenging due to significant appearance changes of the individual subjects caused by aging, facial expressions, glasses, hair, moustache, non-uniform illumination variations and slight changes in pose. Table 2.3 summarizes our results.

Table 2.3: Automatic face recognition experiment in FERET database.

Methods	Recognition accuracy	
	(FA, DupI)	(FA, DupII)
HDF-Fusion	62.6359%	56.1404%
HDF-HOG	72.4185%	**64.9123%**
HDF-SIFT	63.3152%	51.3158%
HDF-LBP	62.5000%	56.1404%
GradientCorr-IC	**75.9511%**	60.5263%

Experiments on MultiPIE database

We also perform extensive experiments on the Multi-PIE dataset to verify the generalization ability of our approach. The Multi-PIE dataset contains face images from 337 subjects, imaged under 15 view points and 19 illumination conditions in 4 recording sessions. Moreover, Multi-PIE is collected under a controlled setting systematically simulating the effects of pose, illumination, and expression. In contrary to the protocols used in the past by the state of the art methods, we decide to perform really challenging experiments by using only one picture per person under illumination 05 as our training set. We have conducted two sets of recognition experiments. In the first set of experiments, we have tested the methods under different face illumination conditions. More specifically, the testing sets consist of:

- one frontal photo per person under illumination 01 (Experiment 1).
- one frontal photo per person under illumination 13 (Experiment 2).
- one frontal photo per person under illumination 19 (Experiment 3).
- one frontal photo per person under illumination 00 (Experiment 4).
- one frontal photo per person under illumination 09 (Experiment 5).
- one frontal photo per person under illumination 15 (Experiment 6).

The face recognition accuracy for the first set of experiments is reported in table 2.4.

In the second set of experiments, we have tested the methods under different facial expressions. More specifically, the testing sets consist of:

- one smile frontal photo per person taken from session 1 under illumination 15 (Experiment 7).
- one squint frontal photo per person taken from session 2 under illumination 15 (Experiment 8).

Table 2.4: Face recognition experiments in MultiPIE database under different face illumination conditions.

Methods	Recognition accuracy					
	Experiment id					
	1	2	3	4	5	6
HDF-Fusion	46.2908%	16.9139%	59.9407%	61.7211%	39.4659%	99.1098%
HDF-HOG	**97.3294%**	**70.9199%**	**99.1098%**	**100%**	95.5490%	**100%**
HDF-SIFT	82.7893%	35.0148%	96.4392%	96.7359%	85.7567%	**100%**
HDF-LBP	45.9941%	16.6172%	59.9407%	61.7211%	39.4659%	99.1098%
GradientCorr-IC	95.2522%	65.5786%	98.8131%	99.4065%	**96.7359%**	**100%**

- one smile frontal photo per person taken from session 3 under illumination 15 (Experiment 9).
- one disgust frontal photo per person taken from session 3 under illumination 15 (Experiment 10).

The face recognition accuracy for the second set of experiments is reported in table 2.5.

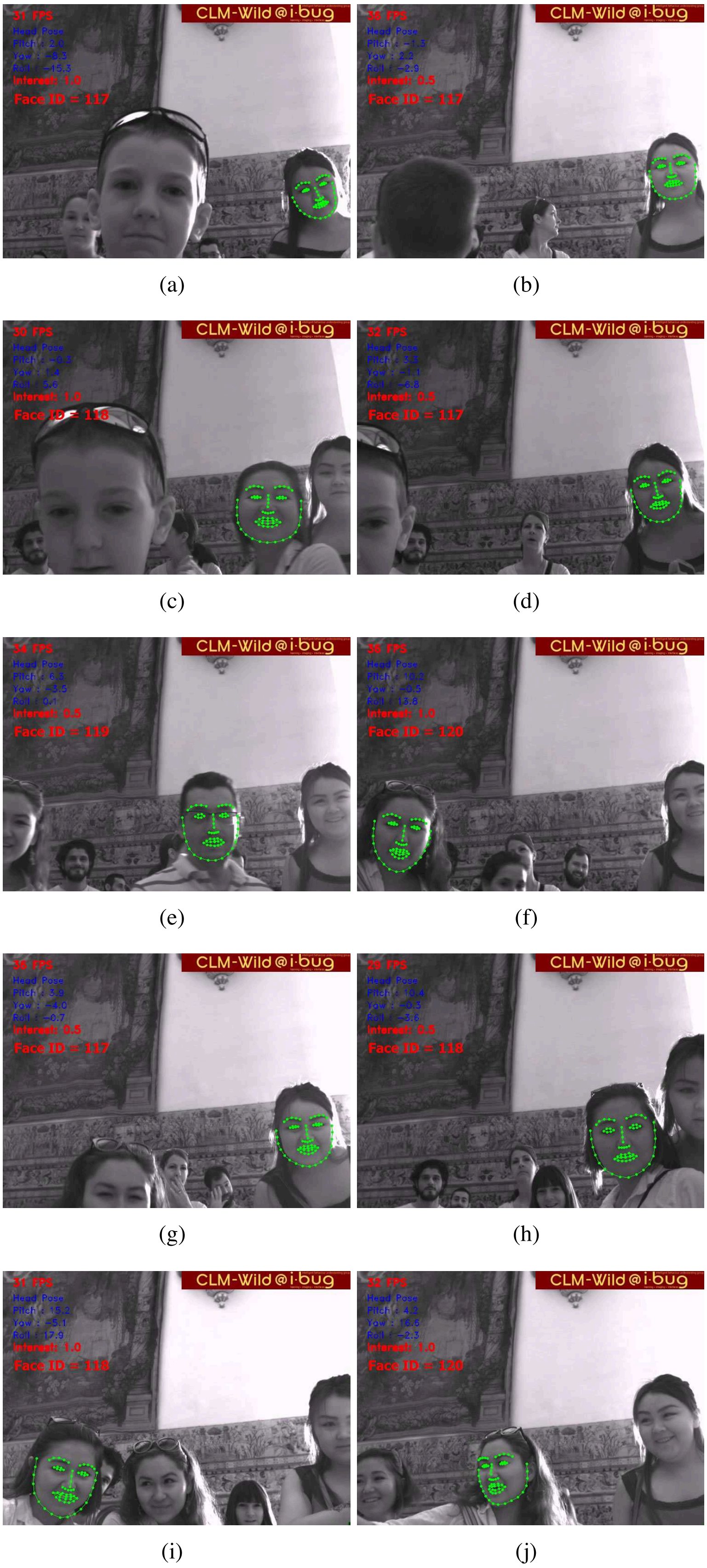
Table 2.5: Face recognition experiments in MultiPIE database under different facial expressions.

Methods	Recognition accuracy			
	Experiment id			
	7	8	9	10
HDF-Fusion	35.7430%	20.6897%	14.7826%	13.9130%
HDF-HOG	74.6988%	48.7685%	38.6957%	35.6522%
HDF-SIFT	79.5181%	53.2020%	45.2174%	41.7391%
HDF-LBP	35.7430%	20.6897%	14.7826%	13.9130%
GradientCorr-IC	**81.9277%**	**83.2512%**	**46.5217%**	**61.3043%**

Based on the experiments in Multi-PIE database it is obvious that GradientCorr-IC performs significantly better than the method in [12] especially when facial expressions used as testing sets. The reason is that that GradientCorr-IC can overcome better the negative effect of working with automatic detected facial landmarks especially in case of facial expressions.

Experiments on FROG data

In FROG project, the robot only has to remember a limited number of faces belonging to the visitors interacted with it since the beginning of the tour. These faces are stored in a non-persistent in-memory database called the gallery. For each frame, the robot compares every visible human face with every face stored in this gallery. For our face recognition experiments in FROG data, we have collected by about 4 hours of real FROG video recordings. In each video, up to 140 different persons were depicted. Our face recognition experiments show that the 99.8% of the human faces can be correctly recognised, which means that our method performs extremely well in outdoor environments. Figure 2.9 depicts some real visual FROG face recognition results.



(a) (b) (c) (d) (e) (f) (g) (h) (i) (j)

Figure 2.9: Real visual FROG face recognition results.

Chapter 3

Spontaneous Agreement and Disagreement Recognition

3.1 Agreement and Disagreement

3.1.1 Definitions and Associated Cues

Distinguishing between different kinds of agreement and disagreement is difficult, mainly because of the lack of a widely accepted definition of (dis)agreement [10]. We can distinguish among at least three ways one could express (dis)agreement with:

- **Direct Speaker's (Dis)Agreement:** A speaker directly expresses his/her (dis)agreement, e.g. "I (dis)agree with you".
- **Indirect Speaker's (Dis)Agreement:** A speaker does not explicitly state her (dis)agreement, but expresses an opinion that is congruent (agreement) or contradictory (disagreement) to an opinion that was expressed earlier in the conversation.
- **Nonverbal Listener's (Dis)Agreement:** A listener nonverbally expresses her (dis)agreement to an opinion that is currently or was just expressed. This could be via auditory cues like "mm hmm" or visual cues like a head nod or a smile.

It is important to mention at this point that in spontaneous direct and indirect speaker's (dis)agreement, the speaker also exhibits nonverbal behavior which could perhaps be different than the one exhibited during the nonverbal listener's (dis)agreement.

Tables 3.1 and 3.2 present a full list of the nonverbal cues that can be displayed during (dis)agreement [10]. The most prevalent and straightforward cues seem to be the **Head Nod** and the **Head Shake** for agreement and disagreement respectively, with nods intuitively conveying affirmation and shakes negation. However, simply the presence of these or any of the other cues alone cannot be discriminative enough, since they could have many other interpretations, as studied by Poggi et al. [44] and Kendon [32].

CUE	**KIND**
Head Nod	Head Gesture
Listener Smile (AU12, AU13) Eyebrow Raise (AU1+AU2)+Head Nod AU1 + AU2 + Smile (AU12, AU13)	Facial Action Facial Action, Head Gesture Facial Action
Sideways Leaning	Body Posture
Laughter Mimicry	Audiovisual Cue Second–order Cue

Table 3.1: Cues of Agreement. For relevant descriptions of AUs, see FACS [17].

3.1.2 Related Work on Automatic Recognition

There is no work, to the best of our knowledge, that has attempted (dis)agreement classification on audiovisual spontaneous data. Table 3.3 summarizes the existing systems that have attempted classification of agreement and/or disagreement in one way or another. However, none of these systems is directly comparable with ours.

Hillard et al. [29] attempted speaker (dis)agreement classification on pre–segmented 'spurts', speech segments by one speaker with pauses not greater than 500ms. The authors used a combination of word–based and prosodic cues to classify each spurt as 'positive–agreement', 'negative–disagreement', 'backchannel', or 'other'. Most of the results reported included word–based cues, however an overall classification accuracy of 62% was reported for a 17% confusion rate between the agreement and disagreement classes. Similar works by Galley et al. [22] and Hahn et al. [27] also deal with classifying spurts as disagreement and agreement, with [22] also dealing with finding the addressee of the action. Germesin and Wilson [24] also deal with these issues. However, the features used by these works included lexical, structural and durational cues and are not comparable with other systems based on non–verbal cues.

The first such system is that by el Kaliouby and Robinson [18], which attempted (dis)agreement classification of *acted* behavioural displays based on head and facial movements. They used 6 classes: 'agreeing', 'disagreeing', 'concentrating', 'interested', 'thinking', and 'unsure'. They tracked 25 fiducial facial points, out of which they extrapolated rigid head motion (yaw, pitch, and roll), and facial action units (eyebrow raise, lip pull, lip pucker), but also utilized appearance–based features to summarise mouth actions (mouth stretch, jaw drop, and lips parting). They used Hidden Markov Models (HMMs) to detect each head and facial action, and a Dynamic Bayesian Network (DBN) per class was trained to perform the higher–level inference of each of the 'mental states' mentioned above, allowing for the co–occurrence of states.

Sheerman–Chase et al. [46] are, to our knowledge, the only research group who have attempted recognition of agreement based on non–verbal cues in spontaneous data. However, they did not include disagreement as a class, because of the lack of data. They instead distinguished between 'thinking', 'understanding', 'agreeing' and 'questioning'. Their spontaneous data was obtained by capturing the four 12–minute dyadic conversa-

CUE	**KIND**
Head Shake	Head Gesture
Head Roll	Head Gesture
Cut Off	Head Gesture
Clenched Fist	Hand Action
Forefinger Raise	Hand Action
Forefinger Wag	Hand Action
Hand Chop	Hand Action
Hand Cross	Hand Action
Hand Wag	Hand Action
Hands Scissor	Hand Action
Ironic Smile/Smirking [AU12 L/R(+AU14)]	Facial Action
Barely noticeable lip–clenching (AU23, AU24)	Facial Action
Cheek Crease (AU14)	Facial Action
Lowered Eyebrow/Frowning (AU4)	Facial Action
Lip Pucker (AU18)	Facial Action
Slightly Parted Lips (AU25)	Facial Action
Mouth Movement (AU25/AU26)	Facial Action
Nose Flare (AU38)	Facial Action
Nose Twist (AU9 L/R, AU10 L/R, AU11 L/R)	Facial Action
Tongue Show (AU19)	Facial Action
Suddenly Narrowed/Slitted Eyes (fast AU7)	Facial Action
Eye Roll	Facial Action/Gaze
Gaze Aversion	Gaze
Arm Folding	Body Posture
Large Body Shift	Body Action
Leg Clamp	Body Posture
Head/Chin Support on Hand	Body/Head Posture
Neck Clamp	Hand/Head Action
Head Scratch	Head/Hand Action
Self–manipulation	Hand/Facial Action
Feet Pointing Away	Feet Posture
Sighing	Auditory Cue
Throat Clearing	Auditory Cue
Delays	Auditory Cue
Utterance Length	Auditory Cue
Interruption	Auditory Cue

Table 3.2: Cues for Disagreement. For relevant descriptions of AUs, see FACS [17]

tions of 6 males and 2 females. 21 annotators rated the clips with each clip getting on average around 4 ratings that were combined to obtain the ground truth label. For the automatic recognition, they used no auditory features and the tracking of 46 fiducial facial points was used. The output of the tracker was then processed to obtain a number of static and dynamic features to be used for classification. Principal Component Analysis (PCA) was performed on the tracked points in each video frame, and the PCA eigenvalues were used as features. Similarly to el Kaliouby and Robinson [18], the head yaw, pitch and roll, the eyebrow raise, lip pucker and lip parting were calculated as functions of these tracked facial points. Gaze was also estimated in a similar fashion —the eye pupils were among the points tracked.

3.2 Hidden conditional random fields (HCRFs) for Multimodal Gesture Recognition

We wish to learn a mapping of observations $\mathbf{x}$ to class labels $y \in Y$, where $\mathbf{x}$ is a vector of m local observations, $\mathbf{x} = \{x_1, x_2, \ldots, x_m\}$, and each local observation x_j is represented by a feature vector $\phi(x_j) \in \Re^d$. An HCRF models the conditional probability of a class label given an observation sequence by:

$$P(y \mid \mathbf{x}, \theta) = \sum_{\mathbf{s}} P(y, \mathbf{h} \mid \mathbf{x}, \theta) = \frac{\sum_{\mathbf{h}} e^{\Psi(y,\mathbf{h},\mathbf{x};\theta)}}{\sum_{y' \in Y, \mathbf{s}} e^{\Psi(y',\mathbf{h},\mathbf{x};\theta)}}. \tag{3.1}$$

where $\mathbf{h} = \{h_1, h_2, \ldots, h_m\}$, each $h_i \in H$ captures certain underlying structure of each class and H is the set of hidden states in the model. The potential function $\Psi(y, \mathbf{h}, \mathbf{x}; \theta) \in \Re$ is an energy function, parameterized by θ, which measures the compatibility between a label, a sequence of observations and a configuration of the hidden states. The following objective function is used in training the parameters:

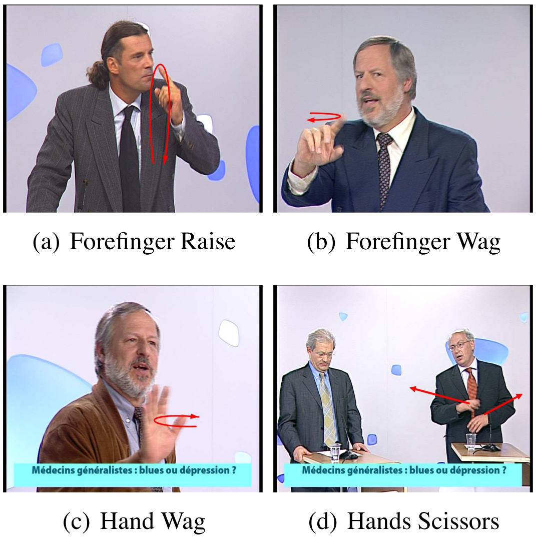
$$L(\theta) = \sum_i \log P(y_i \mid x_i, \theta) - \frac{1}{2\sigma^2}||\theta||^2 \tag{3.2}$$

The first term in Eq. 3.2 is the log-likelihood of the data. The second term is the log of a Gaussian prior with variance σ^2, i.e., $P(\theta) \sim \exp\left(\frac{1}{2\sigma^2}||\theta||^2\right)$. We use gradient ascent to search for the optimal parameter values, $\theta^* = \arg\max_\theta L(\theta)$, under this criterion. For our experiments we used a Quasi-Newton optimization technique to minimize the negative logliklihood of the data.

Hidden conditional random fields (HCRFs) —discriminative models that contain hidden states— are well–suited to the problem of multimodal cue modeling for agreement/disagreement recognition. Quattoni [45] presented and used HCRFs to capture the spatial dependencies between hidden object parts. Wang et al. [52] used them to capture temporal dependencies across frames and recognize different gesture classes. They did so successfully by learning a state distribution among the different gesture classes in a discriminative manner, allowing them to not only uncover the distinctive configurations that uniquely identifies each class, but also to learn a shared common structure among

Method	Features	Classifier	Data	Spontaneous
Method in [29]	Verbal, pause, fundamental, frequency(F0), duration	Decision Tree	ICSI [31]	√
Method in [22]	Verbal	Bayesian Network	ICSI [31]	√
Method in [18]	head nod, head shake, head turn, head tilt, AU1, AU2, AU12, AU16, AU19, AU20, AU25, AU26, AU27	HMM, DBN	Mind Reading DVD [7]	—
Method in [27]	Verbal	Contrast Classifier, SVM	ICSI [31]	√
Method in [46]	head yaw, head pitch, head roll, AU1, AU2, AU12, AU18, AU20, AU25, Gaze, head pose	AdaBoost	own	√
Method in [24]	Verbal, pitch, energy, duration, pauses, speech rate	Decision Tree, CRF	AMI [11]	√

Table 3.3: Summary of the existing systems that have attempted (dis)agreement classification.


Médecins généralistes : blues ou dépression ?
Médecins généralistes : blues ou dépression ?

(a) Forefinger Raise (b) Forefinger Wag

(c) Hand Wag (d) Hands Scissors

Figure 3.1: Some of the gestures used as cues for the experiments.

the classes. Moreover, as a discriminative model, HCRFs require a fewer number of observations than a generative model like a Hidden–Markov Model (HMM). These were all qualities that prompted us to select HCRFs as a model to experiment with in our attempt to recognize (dis)agreement.

Finally, another very important quality is the ability to easily investigate an HCRF model and find out what pieces of information it learned to be most important for the task at hand. This can easily be done by obtaining the weights learned for each of its features and ranking them. In our experiments, the features we used were the actual observation vector, the "edge" feature function between two hidden states, and the "label edge" feature function between a hidden state and a class label. Therefore, for a three-state HCRF, like the one in figure 3.3, the features included the 10 cues as those are listed in Table 3.4, the 9 (3x3) transitions from one hidden state to the other, and the 6 (3x2) transitions from each hidden state to a class label. After the HCRF is trained on a training set, we are able to sort the weights it learned for each set of features and derive useful conclusions about what hidden states are associated with which label, which features are most important in each hidden state, and what are the most probable transitions given a current hidden state.

3.3 Experiments

3.3.1 Dataset and Cues

Our dataset originated from the *Canal 9 Database of Political Debates* [51], one that comprises of 43 hours and 10 minutes of 72 real televised debates on *Canal 9*, a local Swiss television station. The debates are moderated by a presenter, and there are two sides that argue around a central issue, with one or more participants on each side. Hence, the database is rather rich in episodes of spontaneous (dis)agreement.

The dataset we used comprises of 53 episodes of agreement and 94 episodes of disagreement, which occur over a total of 11 debates. These episodes were selected on the basis of the verbal content, and thus, only episodes of direct and indirect (dis)agreement were included (see Sect. 3.1.1). As the debates were filmed with multiple cameras, and edited live to one feed, the episodes selected for the dataset were only the ones that were contained within one personal, close–up shot of the speaker.

We automatically extracted nonverbal auditory features used in related work, specifically the fundamental frequency (F0) and energy, by using *OpenEar*[20]. Since we are interested in answering questions relevant to the automatic recognition of (dis)agreement based on the dynamics of multimodal cues, we manually annotated the dataset for a number of hand and head gestures, in order to gather as accurate temporal information about the gestures as possible. The cues we finally extracted and used in our experiments are listed in Table 3.4; the visual cues that may not be self–explanatory from their title are depicted in figure 3.1. The hand and head gestures we included were based off the relevant list of cues from the Social Psychology literature (see Sect. 3.1.1), with the exception of a number of head and hand gestures that never appeared in the dataset, and the addition of the 'Shoulder Shrug' and the 'Forefinger Raise-Like' gestures. The latter is a 'Forefinger Raise' without an erect index finger.

CUE	**KIND**
Head Nod	Head Gesture
Head Shake	Head Gesture
Forefinger Raise	Hand Action
'Forefinger Raise'–Like	Hand Action
Forefinger Wag	Hand Action
Hand Wag	Hand Action
Hands Scissor	Hand Action
Shoulder Shrug	Body Gesture
Fundamental Frequency (F0)	Auditory Cue
Energy	Auditory Cue

Table 3.4: The list of features we used in our experiments.

3.3.2 Methodology

We conducted experiments with Support Vector Machines (SVMs), as our baseline static classifiers, Hidden–Markov Models (HMMs), the most–commonly used dynamic generative model, and Hidden–State Conditional Random Fields (HCRFs), the dynamic discriminative model we believe is most appropriate for such a task. We conducted different experiments for three groups of cues: only auditory, only visual, and both auditory and visual ones.

Our cues were encoded differently for our static and dynamic classifiers, but the same information was available to all classifiers. For SVMs, the features of each gesture were the start frame and the duration (total number of frames) of the gesture within the segment of interest. For the auditory features we used the mean, standard deviation, and the first, second(median), and third quartiles of each. The later values did not take into account the undefined areas of F0, and all values were scaled from -1 to 1. For the experiments with HMMs and HCRFs, we encoded each gesture in a binary manner (1 if the gesture is activated in a certain frame, 0 otherwise), and used the raw values of our auditory features, normalized per subject.

All our experiments were run in a leave–one–debate–out fashion, i.e. the testing set always comprised of examples from the one debate which was not included in the training and validation sets. The optimal model parameters, i.e. number of hidden states for each test set and number of mixtures of Gaussians for HMMs and the number of hidden states and the regularization factor for HCRFs, were chosen by a three–fold validation on the remaining debates. The HMM and HCRF experiments were run with 10 different random initializations, the best of which was chosen each time during the validation phase (i.e., based on performance on the validation sets). The evaluation metric that we used for all the experiments was the total accuracy in a balanced dataset, i.e. percentage of sequences for which the correct label was predicted in a test set that contains an equal number of agreement and disagreement examples.

3.4 Results and Discussion

Figure 3.2 summarizes the results of the experiments on spontaneous agreement and disagreement classification using auditory, gestural and both auditory and gestural features. It is clear that:

(a) It *is* possible to perform the task of spontaneous agreement and disagreement classification without the use of any verbal features.

(b) The temporal dynamics of the cues are vital to the task, as it is evident that SVMs are not able to perform well by using static information alone.

(c) HCRFs outperform SVMs and HMMs, especially when the cues used are multimodal and the underlying dynamics of the different modalities need to be learned.

Figure 3.3 shows how a high–performing three-state HCRF model is able to successfully discriminate between the two social attitudes. By examination of the weights learned by the HCRF for each of its cues, hidden states, and transitions, we were able to rank, according to importance, the information that the model used. The model assigned one state as prevalent for each of the two classes, and one state as shared between them. It also learned what transitions are more likely given each state and each attitude. In the figure the transitions from each state most associated with each attitude are marked as green and red connections, for agreement and disagreement respectively. Also each state contains its highest ranked features in a descending order of importance. The

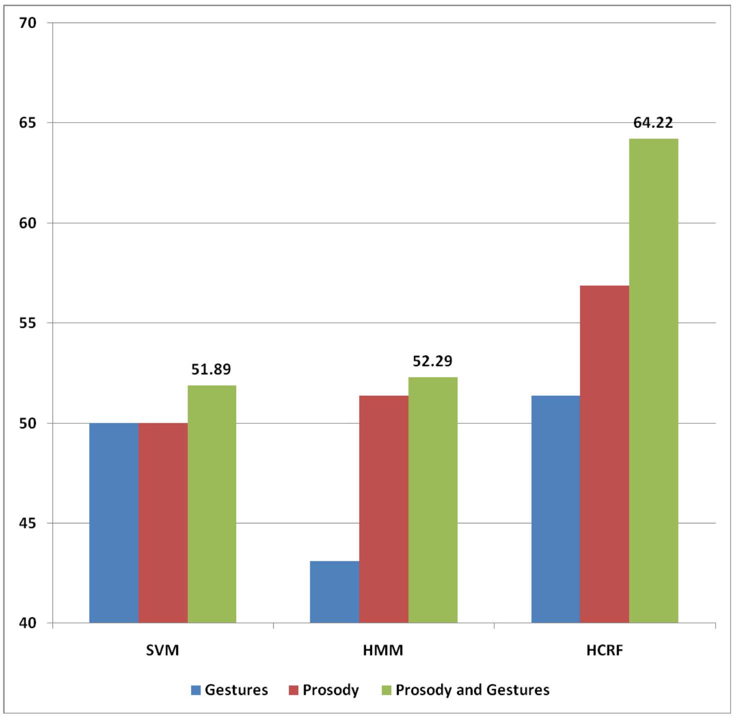


Figure 3.2: Comparisons of recognition performance (total accuracy on a balanced dataset) by the classification methods we explored on the three different groups of features used.

highest ranked features in these 'exclusive' states, combined with the most probable transitions show that the Head Nod and the Head Shake, which are considered, by social psychologists, the most prevalent cues in agreement and disagreement respectively (see Sect. 3.1.1), are also the most discriminative cues here. Finally, it could be the case that 'Forefinger Raise-Like' gestures might in fact play no role in discriminating between the two attitudes.

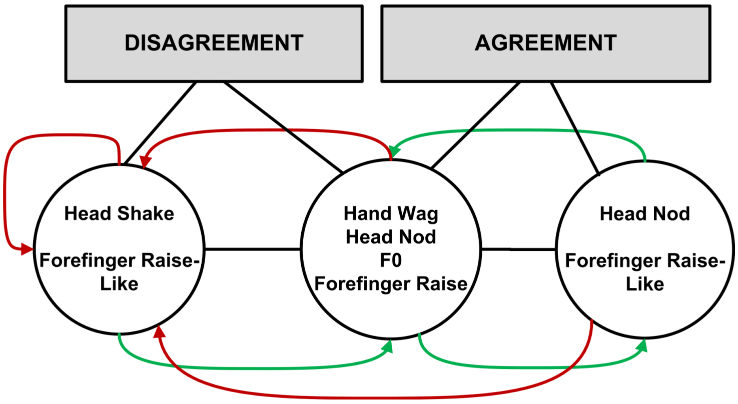
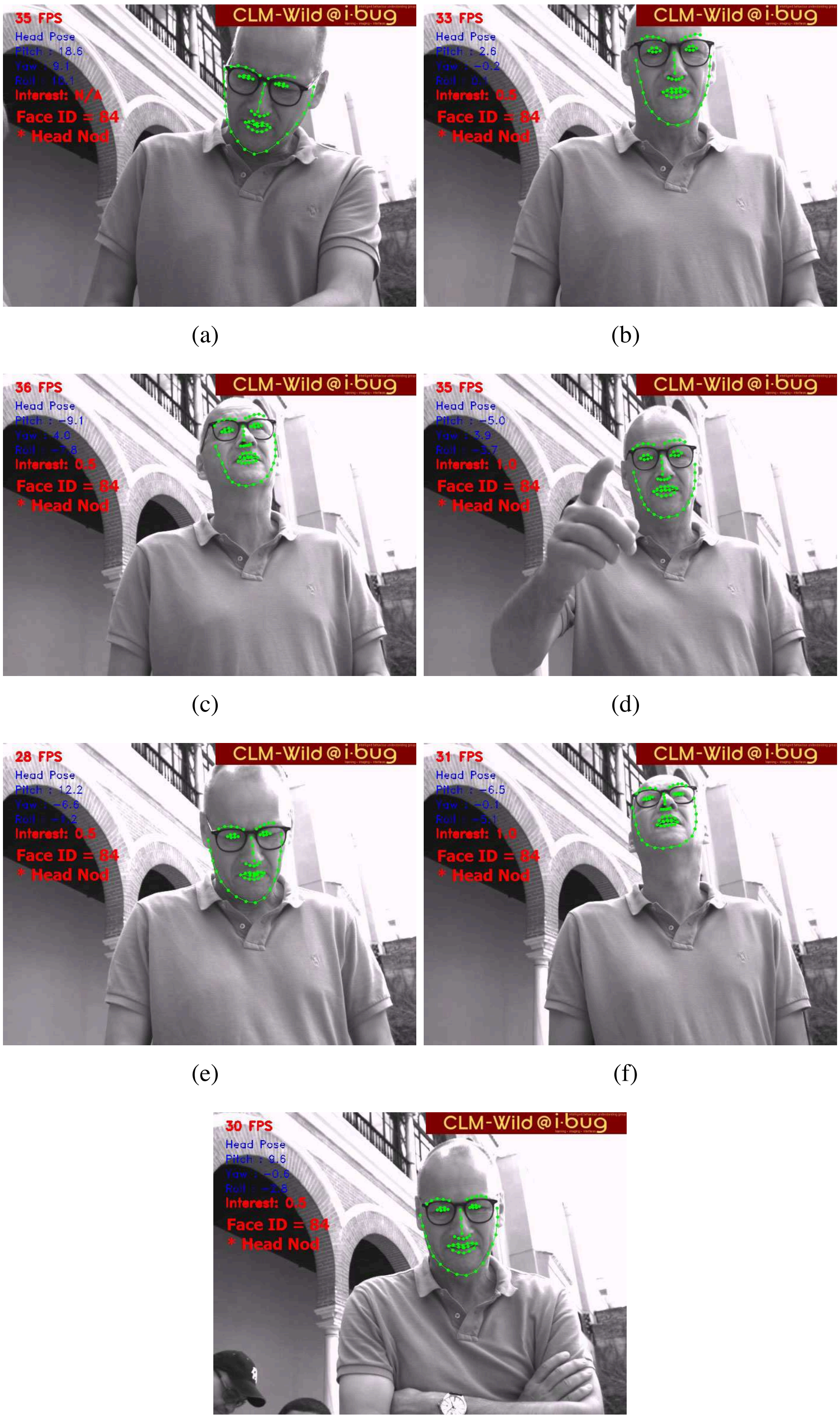


Figure 3.3: The features learned for each state by a three-state HCRF model. The green and red connections correspond to the highest–ranked transition from each state in the cases of agreement and disagreement respectively. The middle state is shared among the two classes.

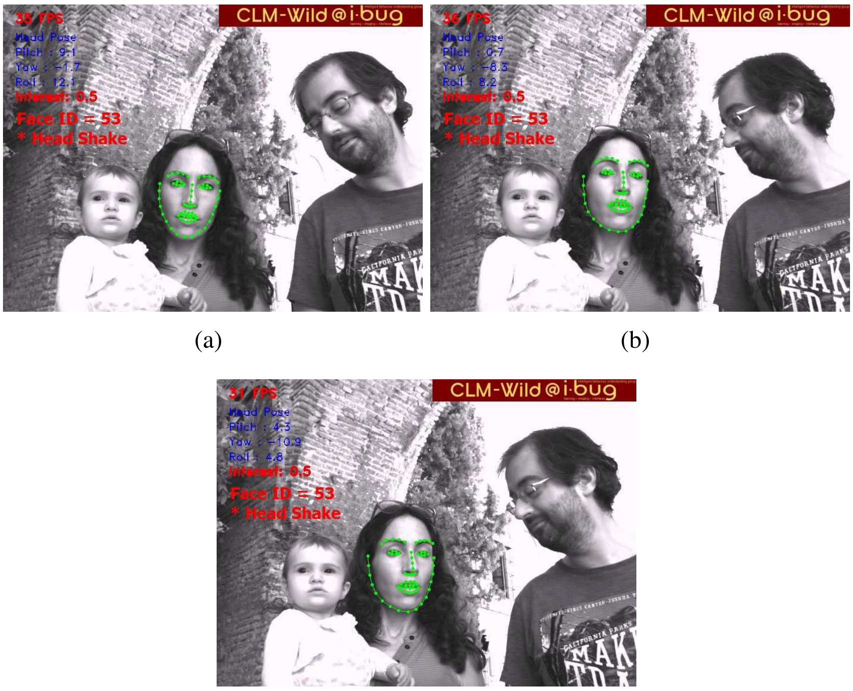
3.4.1 Experiments on FROG data

In FROG project, our framework was used for detection of social attitudes like agreeing and disagreeing. Detection is based on the state of the art in cognitive sciences and based on morphological and temporal correlations between relevant visual cues, including facial gestures like frowns and smiles as well as head gestures like tilts and nods. As features, the facial point landmarks as well as the head pose were used producing very satisfactory results. The method for facial landmark detection as well as for face pose estimation was developed for FROG and it was presented in deliverable 3.1. Figure 3.4 depicts some real visual FROG head nod detection results, while Figure 3.5 depicts some real visual FROG head shake detection results.



(a) (b) (c) (d) (e) (f) (g)

Figure 3.4: Real visual FROG head nod results.



(a)

(b)

(c)

Figure 3.5: Real visual FROG head shake results.

Chapter 4

Conclusion

A method for performing robust human face recognition in uncontrolled (outdoor) environments has been developed for FROG project. Human face recognition is performed in order for the robot to decide if a new visitor comes close to the robot and starts interacting with it. Thus the robot could adapt its tour guide strategy based on the new visitor's implicit affective feedback (e.g. attention and interest). Furthermore, a multi-cue visual method for detection of social attitudes like agreeing and disagreeing has also been developed for FROG project. Detection is based on the state of the art in cognitive sciences and morphological and temporal correlations between relevant visual cues, including facial gestures like frowns and smiles as well as head gestures like tilts and nods. For this purpose, facial landmarks as well as face pose were used. Based on our experiments on real FROG data, this method meets all the requirements of the FROG project regarding the detection of social attitudes like agreeing and disagreeing.

Bibliography

[1] T. Ahonen, A. Hadid, and M. Pietikdinen. Face recognition with local binary patterns. In *Lecture Notes in Computer Science*, pages 469–481, 2004.

[2] A. Ashraf, S. Lucey, and T. Chen. Fast Image Alignment in the Fourier Domain. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 2480–2487, 2010.

[3] S. Baker, R. Gross, T. Ishikawa, and I. Matthews. Lucas-Kanade 20 years on: Part 2. *Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-03-01*, pages 1–47, 2003.

[4] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: Part 3. *Robotics Institute, Carnegie Mellon University, Tech. Rep. CMU-RI-TR-03-35*, pages 1–51, 2003.

[5] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 1090–1097, 2001.

[6] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *International Journal of computer Vision (IJCV)*, 56(3):221–255, 2004.

[7] S. Baron-Cohen, O. Golan, S. Wheelwright, and J. J. Hill. *Mind Reading: The Interactive Guide to Emotions*. London : Jessica Kingsley, 2004.

[8] M. S. Bartlett, J. R. Movellan, and T. J. Sejnowski. Face recognition by independent component analysis. *IEEE Transactions on Neural Network*, 13(6):1450–1464, 2002.

[9] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. fisherfaces: recognition using class specific linear projection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 19(7):711–720, 1997.

[10] K. Bousmalis, M. Mehu, and M. Pantic. Spotting agreement and disagreement: A survey of nonverbal audiovisual cues and tools. In *IEEE International Conference on Affective Computing and Intelligent Interaction*, pages 1–9, 2009.

[11] J. Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation Journal*, 41(2):181–190, 2007.

[12] D. Chen, X. Cao, F. Wen, and J. Sun. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 3025–3032, 2013.

[13] H. Chen, P. Belhumeur, and D. Jacobs. In search of illumination invariants. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 254–261, 2002.

[14] T. Cootes and C. Taylor. On representing edge structure for model matching. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 1114–1119, 2001.

[15] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 886–893, 2005.

[16] N. Dowson and R. Bowden. Mutual information for Lucas-Kanade tracking (milk): An inverse compositional formulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 30(1):180–185, 2007.

[17] P. Ekman, W. V. Friesen, and J. C. Hager. *Facial action coding system*. Salt Lake City: Research Nexus, 2002.

[18] R. el Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 154–154, 2009.

[19] G. Evangelidis and E. Psarakis. Parametric Image Alignment Using Enhanced Correlation Coefficient Maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, pages 1858–1865, 2008.

[20] F. Eyben, M. Wfllmer, and B. Schuller. openear-introducing the munich open-source emotion and affect recognition toolkit. In *IEEE International Conference on Affective Computing and Intelligent Interaction*, pages 1–6, 2009.

[21] A. Fitch, A. Kadyrov, W. Christmas, and J. Kittler. Orientation correlation. In *Proceedings of British Machine Vision Conference (BMVC)*, pages 133–142, 2002.

[22] M. Galley, K. McKeown, J. Hirschberg, and E. Shriberg. Identifying agreement and disagreement in conversational speech: use of bayesian networks to model pragmatic dependencies. In *Conference on Meeting Association for Computational Linguistics*, pages 669–676, 2004.

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

[24] S. Germesin and T. Wilson. Agreement detection in multiparty conversation. In *International Conference on Multimodal Interfaces*, pages 7–14, 2009.

[25] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(12):1080–1093, 2005.

[26] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.

[27] S. Hahn, R. Ladner, and M. Ostendorf. Agreement/disagreement classification: Exploiting unlabeled data using contrast classifiers. In *Conference of the NAACL on Human Language Technology*, pages 53–56, 2006.

[28] X. He, S. Yan, Y. Hu, P. Niyogi, and H. J. Zhang. Face recognition using laplacian faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 27(3):328–340, 2005.

[29] D. Hillard, M. Ostendorf, and E. Shriberg. Detection of agreement vs. disagreement in meetings: training with unlabeled data. In *Conference on North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 34–36, 2003.

[30] D. Hond and L. Spacek. Distinctive descriptions for face processing. In *Proceedings of British Machine Vision Conference (BMVC)*, pages 320–329, 1997.

[31] A. Janin, D. Baron, J. Edwards, D. Ellis, D. Gelbart, N. Morgan, B. Peskin, T. Pfau, and E. Shriberg. Icsi meeting corpus. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2003.

[32] A. Kendon. Some uses of the head shake. *Gesture*, 2(2):147–182, 2002.

[33] S. H. Lin, S. Y. Kung, and L. J. Lin. Face recognition/detection by probabilistic decision-based neural network. *IEEE Transactions on Neural Network*, 8(1):114–132, 1997.

[34] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of computer Vision (IJCV)*, 60(2):91–110, 2004.

[35] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *International joint conference on artificial intelligence*, pages 674–679, 1981.

[36] A. Martinez and R. Benavente. The AR Face Database. CVC Technical Report# 24, 1998.

[37] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of computer Vision (IJCV)*, 60(2):135–164, 2004.

[38] R. Megret, J. Authesserre, and Y. Berthoumieu. Bidirectional Composition on Lie Groups for Gradient-Based Image Alignment. *IEEE Transactions on Image Processing*, 19(9):2369–2381, 2010.

[39] A. Papoulis, S. Pillai, and S. Unnikrishna. *Probability, random variables, and stochastic processes*. McGraw-Hill New York, 2002.

[40] A. Patel and W. Smith. 3d morphable face models revisited. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 1327–1334, 2009.

[41] P. J. Phillips, P. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 947–954, 2005.

[42] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The feret evaluation methodology for face recognition algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, 2000.

[43] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing*, 16(5):295–306, 1998.

[44] I. Poggi, F. D. Errico, and L. Vincze. Types of nods. the polysemy of a social signal. In *International Conference on Language Resources and Evaluation*, 2010.

[45] A. Quattoni, M. Collins, and T. Darrell. Conditional random fields for object recognition. In *Conference on Neural Information Processing Systems*, pages 1097–1104, 2004.

[46] T. Sheerman-Chase, E.-J. Ong, and R. Bowden. Feature selection of facial displays for detection of non verbal communication in natural conversation. In *IEEE International Workshop on Human-Computer Interaction*, 2009.

[47] M. A. Turk and A. P. Pentland. Face recognition using eigenfaces. In *IEEE Conference On Computer Vision and Pattern Recognition (CVPR)*, pages 586–591, 1991.

[48] G. Tzimiropoulos, V. Argyriou, S. Zafeiriou, and T. Stathaki. Robust FFT-Based Scale-Invariant Image Registration with Image Gradients. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 39:1899–1906, 2010.

[49] G. Tzimiropoulos and S. Zafeiriou. On the Subspace of Image Gradient Orientations. *Arxiv preprint arXiv:1005.2715*, 2010.

[50] Y. Ukrainitz and M. Irani. Aligning sequences and actions by maximizing space-time correlations. In *Computer Vision–European Conference on Computer Vision (ECCV)*, pages 538–550, 2006.

[51] A. Vinciarelli, A. Dielmann, S. Favre, and H. Salamin. Canal9: A database of political debates for analysis of social interactions. In *IEEE International Conference on Affective Computing and Intelligent Interfaces*, pages 96–99, 2009.

[52] S. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, and T. Darrell. Hidden conditional random fields for gesture recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1521–1527, 2006.

[53] X. Xiong and F. De la Torre. Supervised descent method and its application to face alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013.