ELSEVIER

Signal Processing: *Image Communication* 16 (2001) 585–608

SIGNAL PROCESSING: IMAGE COMMUNICATION

www.elsevier.nl/locate/image

# A visual analysis/synthesis feedback loop for accurate face tracking

Stéphane Valente[1], Jean-Luc Dugelay*

*Institut Eurécom, Multimedia Communications Department, B.P. 193, 06904 Sophia-Antipolis, France*

Received 7 October 1998

**Abstract**

We propose a novel approach for face tracking, resulting in a visual feedback loop: instead of trying to adapt a more or less realistic artificial face model to an individual, we construct from precise range data a specific texture and wireframe face model, whose realism allows the analysis and synthesis modules to visually cooperate in the image plane, by directly using 2D patterns synthesized by the face model. Unlike other feedback loops found in the literature, we do not explicitly handle the 3D complex geometric data of the face model, to make real-time manipulations possible. Our main contribution is a complete face tracking and pose estimation framework, with few assumptions about the face rigid motion (allowing large rotations out of the image plane), and without marks or makeup on the user's face. Our framework feeds the feature-tracking procedure with synthesized facial patterns, controlled by an extended Kalman filter. Within this framework, we present original and efficient geometric and photometric modelling techniques, and a reformulation of a block-matching algorithm to make it match synthesized patterns with real images, and avoid background areas during the matching. We also offer some numerical evaluations, assessing the validity of our algorithms, and new developments in the context of facial animation. Our face-tracking algorithm may be used to recover the 3D position and orientation of a real face and generate a MPEG-4 animation stream to reproduce the rigid motion of the face with a synthetic face model. It may also serve as a pre-processing step for further facial expression analysis algorithms, since it locates the position of the facial features in the image plane, and gives precise 3D information to take into account the possible coupling between pose and expressions of the analysed facial images. 

*Keywords:* Analysis/synthesis cooperation; Face tracking; Face modelling; Kalman filter; Face cloning; Virtual teleconferencing; MPEG-4

## 1. Context

### 1.1. Facial animation in MPEG-4

*Facial animation* is currently a hot topic for the MPEG-4 standard [29], which created the Synthetic/Natural Hybrid Coding (SNHC) working group [26] to mix real and computer-generated objects for

[1] Present address: "Ecole des Mines de Paris", Centre CAOR, 60, boulevard Saint Michel, 75006 Paris, France.

* Corresponding author.

*E-mail address:* dugelay@eurecom.fr; http://www.eurecom.fr/~image (J.-L. Dugelay).

0923-5965/01/$ - see front matter 
PII: S0923-5965(00)00038-2

the reproduction of real-world scenes. MPEG-4 proposes the use of a synthetic face model in a wide range of applications, from virtual actors, video games, human–machine communication, story tellers on demand, multimedia interactive environments, video-telephony and, as in the context of this work, virtual conferencing [17,19,28,38,40]. Facial animations are controlled by face animation parameters (FAP), which manipulate the displacements and angles of the face's features.

MPEG-4 defines a stream syntax and a decoder for the synthetic face model, but does not provide or impose any technique for obtaining the stream of facial animations, which is where many contributions are still possible. For example, it allows the use of text-to-speech systems for automated talking heads, or hand-defined animations. However, animating a face model given the performance of a real actor or user (*face cloning*), although it is an unsolved problem, remains a major issue, and could ensure the acceptance and success of future services based on this standard.

The main contribution of this paper is a rigid motion estimation algorithm of a face in a video sequence, that achieves near real-time performance, without any constraint on the user (no markers or makeup, unknown lighting and no specific background), allowing large rotations of the head. Offline, our framework can be used to generate precise orientation parameters for a MPEG-4 animation stream (FAP 48 to 50), regardless of the method used to obtain the other FAPs, be it from speech, text and hand. It can be used, for instance, in conjunction with other algorithms extracting facial expressions from images, or the animation rules of automated talking heads to obtain head motions with life-like timings (timings that happen to be difficult to obtain with key-frame animation systems by hand, even for animation experts). Online, apart from MPEG-4 systems, it can also be integrated in a virtual teleconferencing system, recovering the head pose during the session.

### *1.2. Face-cloning for virtual teleconferencing*

Face-cloning techniques are particularly useful for *virtual teleconferencing* systems, which is the context of this work [10]: the key idea is to provide the participants with a common meeting space as if they were all sitting in the same physical room, and to offer them individual points of view, in accordance with the virtual position they occupy in the meeting space. In addition to more flexible visualization possibilities, such as the scene restitution from different points of view [15], the model-based aspect of face cloning allows the meeting to be run over low bit-rate networks, such as the Internet, or mobile networks. However, due to the 3D and interactive nature of the system, several constraints must be added to the face-cloning algorithm:

- the face analysis and synthesis frame-rates, and the image-processing delays, should be as low as possible;
- the synthetic clones can be rendered from any point of view, and a full 3D model is needed (not only the frontal part of the face);
- the face cloning system should operate without colored marks taped on the performer's face;
- it should deal with unknown lighting conditions and background;
- the motions and rotations of the user should not be restricted in front of the camera;
- finally, the clones should be visually realistic.

Several efficient 2D face-tracking algorithms exist in the literature, such as [14]. Nevertheless, when large motions are allowed by a face-cloning system, tracking the user's face in the video sequence without recovering the exact position and orientation of the real face (i.e. its global motion) is pointless, firstly because facial expression analysis algorithms need to know the precise location of the facial features in the 2D image, and secondly because some coupling occurs between the pose and the facial expressions: for instance, when the performer looks downward with a neutral expression, his mouth shape will change in the image plane, and the system is likely to misinterpret it as a smile. Being aware of the subject's pose may help the system solve this coupling. Therefore, 2D face-tracking techniques, like color segmentation, hidden markov models, deformable templates, FFT, etc. are more suitable for face recognition or lipreading applications [30].

Assuming that a face model is available (be it generic or person-dependent), we will now concentrate on the different image-processing approaches that allow the 3D rigid motion of the user to be estimated, without any marker, voluntarily omitting 2D-tracking algorithms.

### 1.3. Face tracking and global motion estimation

One method to estimate the rigid motion of an object is to use some kind of motion information, like the optical flow, and interpret it using some shape model, even a simple one. Azarbayejani et al. [3] use feature point tracking, and project the 2D points onto an ellipsoid to incorporate some 3D information for the rigid motion estimation. Basu et al. [4] regularize the velocity field of the user's face with a 3D ellipsoidal model to determine the best fit between the position and orientation of the model and the motion of the image pixels. DeCarlo and Metaxas [8] initialize a polygonal head model on the user's face, and also regularize the optical flow in terms of model displacements. Because their model has a shape closer to the real face than the ellipsoid, their motion estimation is finer. The main problem with these approaches is that they are "blind", in the sense that they have no way to ground the model to the real face, and errors in the estimation accumulate, resulting in a lack of accuracy or tracking failures.

To prevent cumulative estimation errors, an analysis-by-synthesis feedback loop can be introduced at the encoder, adding some texture information to the shape of the face model. Koch [21] extracts some texture from the first video frame to be mapped onto a face model. Images are then synthesized from his parametric scene description, and compared with the original input images of the camera to solve the model's motion. Li et al. [22] use the CANDIDE model and the texture from the first image to iterate the motion estimation process until the synthesis error becomes satisfactory. The strength of their algorithm is that no costly optical flow is computed, instead, they have a direct computation approach, based on an aggregation of data, which is more robust to noise. Eisert and Girod [13] perform the motion estimation on the optical flow between the synthesized image of a realistic textured face model and the real image. In [12], they add an illumination model to prevent breaking the brightness constancy assumption of their optical flow computation. Although these algorithms lead to simple formulations and fewer cumulative errors, interpreting the pixel differences or the pixel velocity field can be costly, because of the linearization of the model 3D displacements in the 2D image frame for each iteration. As a result, they do not claim to be real-time. Moreover, in principle, such algorithms can deal with self-occlusions due to large rigid motion, and the coupling between the pose and expression of the real face, but none of them report their ability to track large head rotations.

Another approach is taken by Schödl et al. [34]. The first image of the face is projected on a generic polygonal face model, and tracking is achieved by synthesizing the face model, and computing the error with the real images. The derivative of the error with respect to the motion parameters is mapped to the intensity gradients in the image, and a steepest descent is iterated until a local minimum of the error is reached. This is also a costly procedure, because the face model must be rendered at each iteration, and changes of the face lighting, combined with the lack of precision of the shape of their model, may lead to an incorrect registration of the synthetic image.

### 1.4. A novel approach based on visual realism

We noticed that in the literature, none of the face-cloning or face-tracking approaches takes advantage of the *visual* realism of their face model to track and/or analyze facial deformations in a more direct manner, as they all rely on the explicit use of all the geometric information contained in the face model to perform the motion regularization, even at the penalizing cost implied by the linearization of the motion of the face vertices in the 2D image plane. We believe that a better alternative to reach near real-time performance, is to use a realistic face model, let computer graphics hardware translate the information of the 3D shape (including self-occlusions and changes of lighting) into the 2D image plane, and work only at the image level,

using little 3D information from the model. Confirming the idea that a dense motion estimation is not necessary, Jebara and Pentland [18] track a few facial features, which are extracted from the first video frame, with correlation-based trackers, and a Kalman filter estimates the 3D motion and structure of the facial feature points. Their system works in real-time, and allows large head rotations. Because they do not have a face model, their algorithm cannot model self-occlusions or changes of lighting due to the face pose, resulting sometimes in incorrect registrations. To overcome these limitations, we propose to update the tracked patterns with synthesized images of the facial feature.

Therefore, we introduce a novel approach for a visual feedback loop, making the analysis and synthesis modules cooperate only at the image level, without explicitly handling the 3D geometric components of the face model within the loop. Our point in this paper is to provide key elements on the advances permitted by such a feedback, in particular to show how a realistic and precise face model, along with appropriate photometric modelling techniques, can be used to robustly track a real face, and accurately recover its rigid motion, with an uncalibrated camera, no markers, unknown lighting conditions, and a non-uniform background.

Section 2 outlines our rigid motion estimation algorithm, as shown in Fig. 1. The next sections describe the technical issues that allow our analysis/synthesis cooperation in the image plane. In Section 3, we present a reconstruction technique that creates realistic face geometric models suitable for real-time synthesis. Section 4 introduces a novel and efficient 3D illumination compensation algorithm while Section 5 details the extended Kalman filter that drives the feedback loop. Section 6 describes the reformulation of the block-matching algorithm to track synthetic patterns in real images, including the robustness to the speaker's background. We discuss in Section 7 the performance of our face-tracking algorithm on a synthetic sequence. Finally, the local animation possibilities of our models are given as perspectives in Section 8, as we are working on the extraction of facial expression parameters from images, still using the visual appearance of our face model.

## 2. Outline of the tracking algorithm

We have developed a face tracking and pose determination system which relies heavily on the cooperation between analysis and synthesis techniques in the image plane. The algorithm can be outlined as follows (see Fig. 1, where the next numbered steps are reported):

**Initialization:**

(a) the user aligns his/her head with the head model, or modifies the initial pose parameters to align the head model with his head. This first alignment obviously requires some user intervention, but it has to be performed only once per session, and we do not consider it as a major issue at the moment. Indeed, some publications have proposed interesting automatic solutions, such as frame-fitting [1,2], or eigenfeatures [18];

(b) when completed, a 3D illumination compensation algorithm is run, to estimate the lighting parameters that will reduce the photometric differences between the synthetic face model and the real face in the user's environment;

**Main loop:**

(i) a Kalman filter predicts the head 3D position and orientation for time $t$;

(ii) the synthetic face model is used to generate an approximation of the way the real face will appear in the video frame at time $t$; this approximation includes both geometric distortions, scales and shaded lighting due to the speaker's pose; the black background of this rendering is used by the encoder as visual clues about the location of the unknown environment with respect to the face facial features, and it will not be used for the 3D parameters restitution;

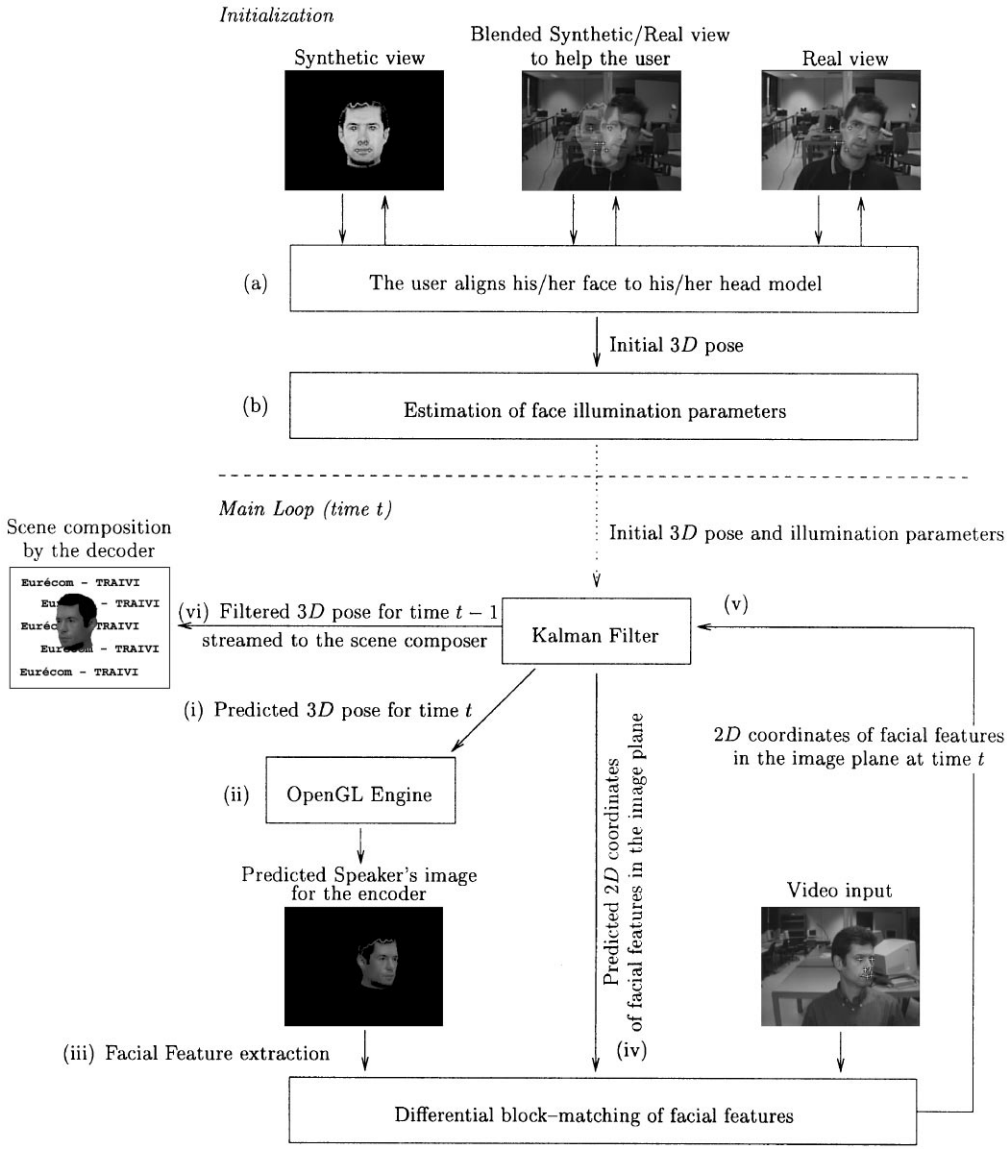


Fig. 1. System overview: initialization, and face-tracking loop – the numbered items (a), (b) and (i)–(vi) are detailed in Section 2.

(iii) patterns are extracted from the synthesized image, representing contrasted facial features (like the eyes, eyebrows, mouth corners, nostrils);
(iv) a differential block-matching algorithm matches these patterns with the user's facial features in the real video frame;
(v) the 2D coordinates of the found positions are given to the Kalman filter, which estimates the current head 3D position and orientation;
(vi) finally, the filtered parameters are quantified, compressed and streamed to the visualization entities, and the rigid motion of the face is recomposed by a decoder, possibly from a different point of view and a new background, or even using a totally different face model.

The strength of the visual feedback loop is that it implicitly takes into account the changes of scale, geometry, lighting and background with almost no overload for the feature-matching algorithm: due to the synthesis module that performs a 3D illumination compensation scheme, the synthesized patterns will predict the geometric deformations, the lighting and the background of the user's facial features, making the differential block-matching pass more robust. This enhanced analysis/synthesis cooperation results in a stable face-tracking framework without artificial marks highlighting the facial features, supports very large rotations out of the image plane (see Fig. 11), and even copes with low-contrasting lightings (see the video demo [23]).

It is clear that such a feedback loop is successful because it combines a realistic face model, efficient lighting modelling techniques, a carefully designed Kalman filter to make up for the uncalibrated camera, and a block-matching algorithm that handles the background interferences during large head rotations. The next sections will deal with these points in details.

## 3. Geometric modelling

We are currently using range data obtained from cylindrical geometry Cyberware range finders [7] to build person-dependent realistic face models. This part is largely independent from the global motion tracking algorithm, and in fact, any realistic face model could be integrated in the analysis/synthesis feedback loop, as long as it is compatible with real-time visualization rates. We should mention here that other techniques are available in the literature to obtain a textured wireframe, using structured lighting for example [31]. In the next paragraphs, we will describe a face model construction algorithm from range data, which provides a high degree of realism while at the same time limitting the number of 3D primitives due to an interactive refinement procedure.

### *3.1. Initial data*

Cyberware scanners produce a dense range image with its corresponding color texture (see Fig. 2(a) and (b)). This dataset is a highly realistic representation of the speaker's face. However, it cannot be used directly in a face cloning system for several reasons. First, this dataset is very dense (an average of 1.4 million vertices) and therefore is not well suited for real-time computations. Furthermore, due to the limitation of the acquisition technology, the dataset is often incomplete and sometimes includes some outliers (as in Fig. 2(a)). Finally, it is not suitable for local deformation computations: it is just a 3D surface, with no anatomical knowledge or built-in animations for the facial features.

### *3.2. Mesh construction*

To achieve both visual realism and real-time computation, we need a geometric model with a limited number of vertices but with enough details in order to distinguish facial features such as the lips or the eyebrows. We use a reconstruction system based on deformable simplex meshes [9] to build such models. Unlike classic approaches, those deformable models are handled as discrete meshes, not relying on any parameterization. Because they are topologically equivalent to triangulations, they can be easily converted as a set of triangles for display purposes or standard 3D file formats like VRML [41]. Finally, they can represent geometric models independently of their topology and they lead to rapid computations.

The different stages of construction from a Cyberware dataset, where the hair information is missing and with some outliers, are shown in Fig. 2. The deformable model is initialized as a sphere (Fig. 2(b)) and then deformed to roughly approximate the face geometry (Fig. 2(c)). The last stage consists in refining the mesh model based on the distance between the range image and the surface curvature (Fig. 2(d)), by selecting the areas where more modelling precision is desired.

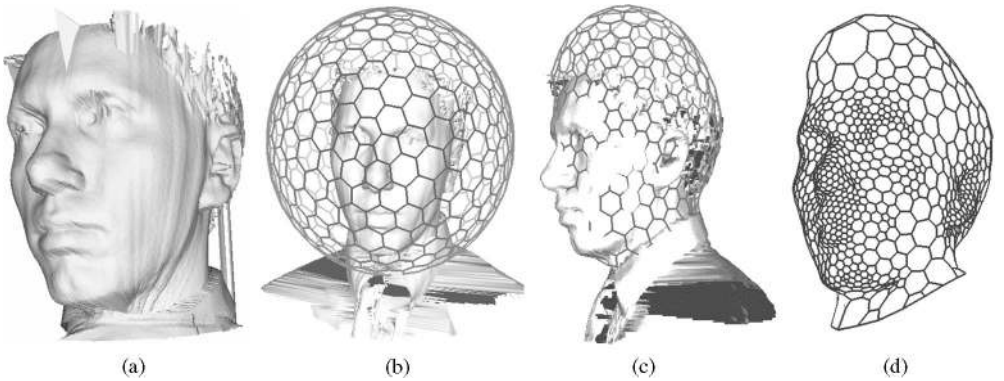


Fig. 2. Reconstruction of a geometric model from a Cyberware dataset: (a) initial range data; (b) initialization; (c) main deformation; (d) mesh refinement – we have interactively selected the areas of interest (chin, ears, nose, lips) where the refinements are performed. The resulting mesh has 2084 vertices.

The face model is then texture-mapped by computing for each vertex of the simplex mesh, its $(u, v)$ coordinates in the range texture image. Since the deformation of each vertex requires the computation of its closest point in the range image, we simply use the texture coordinates of the closest point. Where no range data is available (at the hair level for instance), we project the vertex onto the image plane through the cylindrical transformation of the Cyberware acquisition. This algorithm therefore produces an accurate geometric and texture face model, and has to be run only once for a given head model.

Although some authors reported that adapting a generic face model (CANDIDE) is quite feasible [32,37], we believe that their model could not suit our analysis/synthesis feedback loop due to a low number of primitives, mainly because it would not provide enough lighting normals, which are crucial for the illumination compensation procedure, discussed in the next section.

## 4. Photometric modelling

### 4.1. Motivation

The goal of photometric modelling is to reduce the photometric discrepancies between the speaker's face in the real world environment and his synthetic model directly at the 3D level within the visual feedback loop – otherwise, the block-matching algorithm would constantly fail to match the tracked facial features, because it assumes that the brightness distribution and the shading of synthetic patterns are close to the real facial features, which is far from being true by default (see Fig. 4(a) and (b)). This modelling can be seen as an alternative and elegant technique to other 2D view-based techniques, such as histogram fitting [18]. In [12], Eisert and Girod propose an algorithm to recover the 3D position and intensity of a single infinite light source from a static view assuming an initial guess of the position prior to the motion estimation. Bozdağı et al. [6] have a more complex approach that determines the mean illumination direction and surface albedo to be included in their *optical flow* equation for motion estimation. Both approaches are based on a Lambertian illumination model (i.e. composed of ambient and diffuse lighting) without specular reflections and cast shadows. However, in the real world, cast shadows, and specular highlights (if the user does not have make-up), are likely to occur on a face, and will be difficult to compensate using only a single light as in the previous algorithms.

In [5], Belhumeur and Kriegmann derive that the set of images of a convex Lambertian object under all possible lighting conditions is a cone, which can be constructed from three properly chosen images, and empirically show that cast shadows and specular reflections generally do not damage the conic aspect of the set.

Motivated by the reconstruction possibility of an arbitrary illuminated view from several object images, we propose to recover the face illumination from a single speaker's view by using a set of light sources at different infinite positions. The main advantage of our algorithm is that it can rely on graphics hardware acceleration and compensate unknown light sources with ambient, diffuse and specular components at the 3D level in real-time. A similar idea, applied to interior design, can be found in [35], where the scene's global lighting is computed from the illumination of some objects painted by hand by the scene designer. In our algorithm, the synthetic scene lighting is adjusted by observing the illumination of the facial features in their real environment.

### 4.2. Proposed algorithm

We propose to adopt the following general lighting model, including ambient, diffuse and specular reflections induced by $N$ independent infinite light sources for a 3D textured primitive, with an additional degree of freedom (a luminance offset $L_{\text{offset}}$, which will be justified in Section 4.5):

$$L_{\text{real}} = L_{\text{offset}} + L_{\text{texture}} \times \left( A + \sum_{i=0}^{i=N-1} [(\max\{\boldsymbol{l}_i \cdot \boldsymbol{n}, 0\})D_i + (\max\{\boldsymbol{s}_i \cdot \boldsymbol{n}, 0\})^{\text{shininess}} \times S_i] \right), \quad (1)$$

where $L_{\text{real}}$ denotes the luminance of a pixel in the real image, $L_{\text{texture}}$ the corresponding texture luminance of the face model, $A$ the global ambient light intensity, $D_i$ and $S_i$ the diffuse and specular intensity for the $i$th light, $\boldsymbol{n}$ and $\boldsymbol{l}_i$ the object normal and the $i$th light source direction, $\boldsymbol{s}_i$ the normalized bisector between the $i$th light source direction and the viewing direction, and finally "shininess", the specular exponent controlling the size and brightness of specular highlights.

One can readily verify that the rendered image pixels values in Eq. (1) are linear with respect to the components of the light sources. All the unknowns (the light source intensities, and the luminance offset if needed) can be estimated by a simple least mean square inversion for all the face pixels. Our algorithm thus consists of the following steps:

- manually align the synthetic model with the speaker's image;
- extract, from the real speaker's image, pixel luminance values around the facial features of interest. Over-bright pixels are discarded to avoid areas where the camera sensor might have saturated (the luminance of such pixels would not depend linearly on the contributions of the light sources);
- extract, from the synthetic image, the corresponding texture luminance values $L_{\text{texture}}$ and object lighting normals $\boldsymbol{n}$;
- the intensities $A$, $D_i$, $S_i$ and the offset $L_{\text{offset}}$ are finally estimated by solving Eq. (1) in the least mean square sense.

### 4.3. Numerical evaluation on synthetic images

To validate the assumption that unknown light sources can be compensated by a set of lights at predefined positions, we conducted experiments on synthetic images in order to avoid problems of misalignment between a face model and an unknown image. To that extent, four images of the same model were created respectively with a left diffuse illumination (3(a)), an ambient and left diffuse lighting (Fig. 3(b)), ambient, left diffuse and specular components (3(c)), and ambient, diffuse and specular illuminations from two different light sources (3(d)).

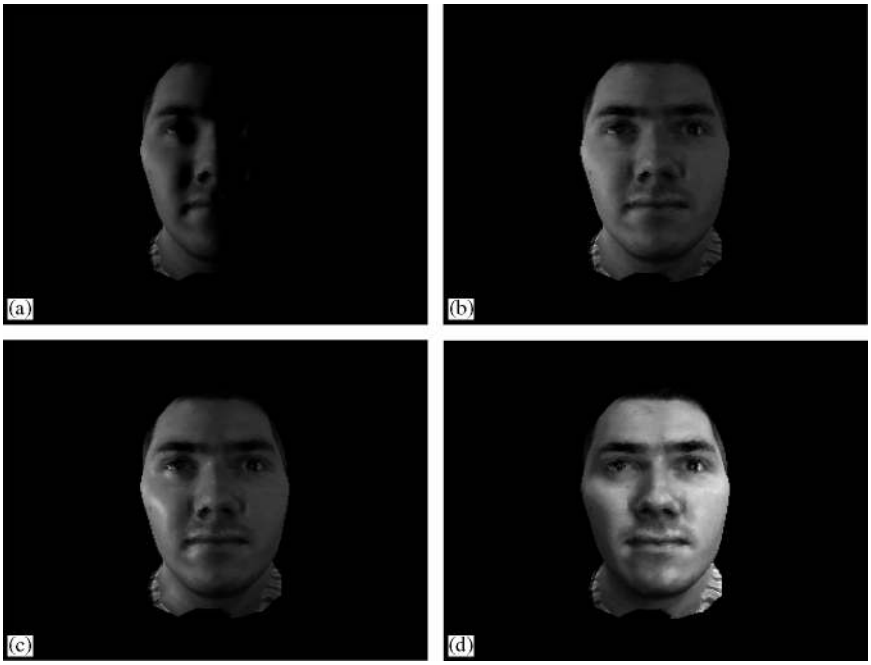


Fig. 3. The synthetic test images used in the validation experiments presented in Table 1. The reconstructed images are not displayed since no significant differences are visible – (a) left diffuse illumination; (b) ambient and left diffuse lighting; (c) ambient, left diffuse and specular lighting; (d) ambient, diffuse and specular components from two light sources.

With these images, we performed two different experiments, *A* and *B*:

(A) we compensated the face illumination with lights located *at the same positions* as the sources used to synthesize the images, enabling and disabling the luminance degree of freedom of Eq. (1). Such experiments can point out numerical differences between the lighting model of Eq. (1), implemented by floating point computations in our software, and the OpenGL lighting operations [27], implemented by dedicated hardware;

(B) we then tried to measure the quality of the compensation using light sources that are *at completely different locations* from the original ones, to evaluate how well unknown light sources can be simulated by lights at arbitrary positions: we compensated for the face illumination with all the light sources on except the ones used for the image creation (our software has seven predefined lights, namely top, bottom, left, right, and three lights around the camera).

Table 1 presents the mean error and the root of the mean square error around the model facial features. The conclusion is twofold:

- with synthetic lights at the right positions, experiments *A* prove that numerical errors are marginal (but exist) in the algorithm, and that Eq. (1) can be correctly implemented by OpenGL;
- experiments *B* suggest that it is fairly reasonable to expect to compensate for ambient, diffuse and specular reflections of unknown intensities and unknown directions by a limited set of lights at predefined locations, without trying to recover the lighting directions.

### 4.4. Experimental results on real images

Fig. 4 shows the illumination compensation for *a real world case*. It is clear that without the compensation (Fig. 4(a)), it would become difficult for a block-matching algorithm to match the synthetic facial features

Table 1
Illumination compensation errors (root of mean square error and mean error in brackets) on synthetic images, for experiments *A* (compensation using the same lighting directions as the original image) and *B* (compensation using all except the same lighting directions as the original image)

| Synthetic images | Fig. 3(a) | Fig. 3(b) | Fig. 3(c) | Fig. 3(d) |
|---|---|---|---|---|
| Without compensation | 77.66 [74.25] | 87.31 [79.92] | 86.79 [78.84] | 44.27 [36.21] |
| Exp. *A*: lights at known positions | 0.24 [0.06] | 0.15 [ − 0.02] | 0.15 [ − 0.02] | 0.36 [0.00] |
| Exp. *B*: lights at unknown positions | 0.30 [0.02] | 4.12 [ − 0.20] | 3.58 [ − 0.43] | 3.56 [ − 0.14] |

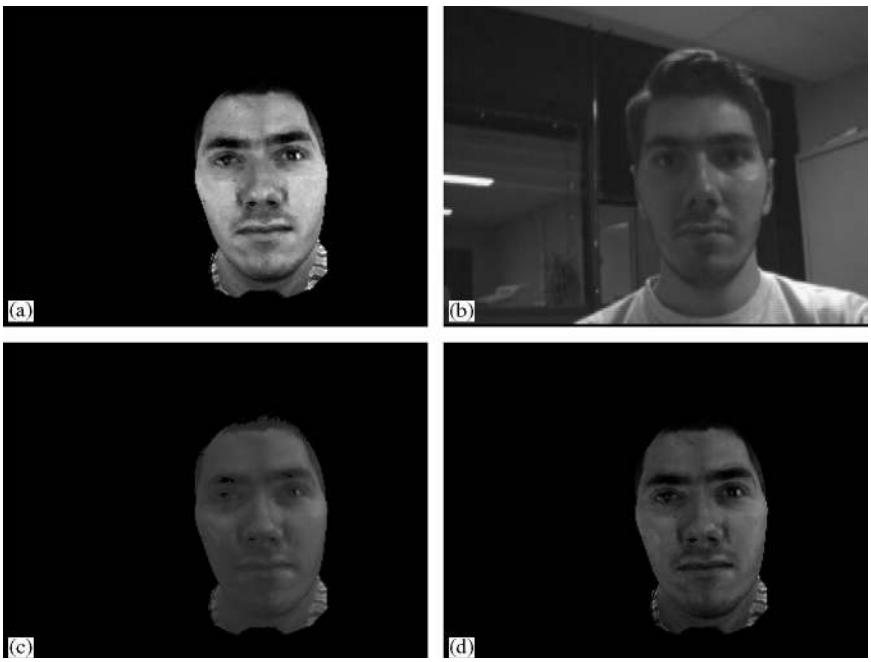


Fig. 4. Illumination compensation on a real face: (a) the speaker's head model with no directional light source; (b) the speaker in a real environment; and the same model with illumination compensation ((c) with and (d) without the illumination offset $L_{\mathrm{offset}}$).

with the real ones (Fig. 4(b)), not only because of the low brightness of the real image, but also because of the shading of the face.

Table 2 displays the numerical errors, before and after the compensation. Our algorithm greatly minimizes the gap between the synthetic and real images (Fig. 4(c) and (b)) by reducing the root of the mean square error by a factor of 6.7, although the enhancements are not as good as for the experiments on synthetic images only. We believe that this is mainly due to the misalignment between the face and the synthetic model, which cannot be perfectly matched "by hand", to the uncalibrated acquisition camera, which does not realize the same perfect perspective projection as the one implemented in our synthesis module, and finally to the texture map of the face model: this texture map should actually correspond to the user's face viewed in ambient lighting, but the scanning device has built-in bright light sources which create parasitic diffuse and specular reflections on the user's face during the texture acquisition. These unpredictible differences are the reason why the brightness offset $L_{\mathrm{offset}}$ was introduced in Eq. (1).

Table 2
Illumination compensation error (root of mean square error and mean error in brackets) between a real and synthetic image

| Real image | (Fig. 4(b)) | |
|---|---|---|
| Without compensation | (Fig. 4(a)) | 62.42 [51.40] |
| With compensation | (Fig. 4(c)) | 9.31 [ − 0.02] |

Table 3
Illumination compensation errors (root of mean square error and mean error in brackets) on synthetic images, without using the illumination offset. These results should be compared to Table 1

| Synthetic images | Fig. 3(a) | Fig. 3(b) | Fig. 3(c) | Fig. 3(d) |
|---|---|---|---|---|
| No compensation | 77.66 [74.25] | 87.31 [79.92] | 86.79 [78.84] | 44.27 [36.21] |
| Exp. *A*: known positions, without $L_{\text{offset}}$ | 0.11 [ − 0.01] | 0.09 [ − 0.01] | 0.09 [ − 0.01] | 0.36 [ − 0.04] |
| Exp. *B*: unknown positions, without $L_{\text{offset}}$ | 0.39 [0.15] | 4.12 [0.04] | 3.56 [ − 0.02] | 3.55 [ − 0.02] |

Table 4
Illumination compensation error (root of mean square error and mean error in brackets) between a real and synthetic image, without the illumination offset. These results should be compared to Table 2

| Real image | (Fig. 4(b)) | |
|---|---|---|
| Without compensation | (Fig. 4(a)) | 62.42 [51.40] |
| With compensation, no $L_{\text{offset}}$ | (Fig. 4(d)) | 15.12 [ − 3.98] |

### 4.5. Justification of $L_{\text{offset}}$

We removed the term $L_{\text{offset}}$ from the lighting model, and reran the compensation algorithm. In Table 3, we see that removing the offset does not bring any significant degradation of the error on synthetic images, which means that Eq. (1) without $L_{\text{offset}}$ corresponds to the lighting operations made by OpenGL to render the synthetic test images.

However, for the compensation on a real image, removing $L_{\text{offset}}$ introduces a higher error (see Table 4), which can be seen in Fig. 4(d), leading to a face model more constrasted than the real face. In this case, the root of the mean square error is only reduced by a factor of 4.1. $L_{\text{offset}}$ is therefore an appropriate degree of freedom, which can be seen as an additional compensation for a difference in mean illumination level between the real and synthetic views, making up for the non-ambient texture of the face model and imprecisions due to the face model's misalignment and the uncalibrated camera.

### 4.6. Concluding remarks about the illumination compensation

We do not claim that the proposed algorithm recovers the exact illumination of the scene, but it helps smooth out the photometric discrepancies around the facial features (not the whole face) between the feedback loop and the video input. Fig. 4 shows that the distributions of dark and bright areas on the real and synthetic views are coherent. The main advantage of our algorithm is that the estimated lighting intensities can be directly injected at the 3D level within the face synthesis module, taking advantage of hardware accelerations that are more and more common on entry-level graphics boards. Furthermore, the lighting

parameters estimation is straightforward and not CPU-demanding, since it does not need to be constrained to output positive light intensities.[2]

To set up the light positions in practice, we found a convenient solution by defining an interface where the user can switch on and off various light sources at predefined positions like the ceiling, the camera, on the left and right hand side… according to his/her environment. As a conclusion, even if it requires some user intervention, we do not think it damages the face-tracking system usability, since it has to be done only once at the beginning of the session. Furthermore, it contributes to the realism of the model (with respect to the speaker's real view), and therefore makes the tracking of real facial features from synthetic patterns by differential block-matching feasible, as described in the next sections.

## 5. Kalman filter

Kalman filters [20] are often used in head tracking systems for two different purposes: the first one is to temporally smooth out the estimated head global parameters, as in [33], the second one is to convert the 2D facial features positions observations into 3D estimates and predictions of the head position and orientation [3]. In our application, the Kalman filter is the central node of our face-tracking system, since it has three different objectives: it recovers the head global position and orientation, it predicts the 2D positions of the feature points for the block-matching algorithm, and – this point is new – it makes the synthesized model have the same scale, position and orientation as the speaker's face in the real view, despite the acquisition by an uncalibrated camera.

### 5.1. Extended Kalman filtering theory

Let us first define a few notations. What we want to estimate is the state vector $\boldsymbol{\Psi}_t$ of a system at time $t$. $\boldsymbol{\Psi}_t$ cannot be measured directly. Instead, we have access to measurements $\boldsymbol{s}_t$ that depend non-linearly on the state vector by the relation $\boldsymbol{s}_t = \boldsymbol{h}(\boldsymbol{\Psi}_t)$, and we have a general idea about the evolution of the system by $\boldsymbol{\Psi}_{t+1} = \boldsymbol{a}(\boldsymbol{\Psi}_t)$. To reflect the uncertainty of the observations and the dynamic evolution of the system, two Gaussian white noises are introduced, denoted $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$, of covariance matrices $\boldsymbol{R}$ and $\boldsymbol{Q}$, and we have the system observation and evolution equations:

$$\boldsymbol{s}_t = \boldsymbol{h}(\boldsymbol{\Psi}_t) + \boldsymbol{v}_t, \qquad \boldsymbol{\Psi}_{t+1} = \boldsymbol{a}(\boldsymbol{\Psi}_t) + \boldsymbol{w}_t. \tag{2}$$

The measurement and evolution functions are linearized respectively around the a priori estimate $\boldsymbol{\Psi}_{t/t-1}$ and the a posteriori estimate $\boldsymbol{\Psi}_{t/t}$ with

$$\boldsymbol{h}(\boldsymbol{\Psi}_t) \approx \boldsymbol{h}(\boldsymbol{\Psi}_{t/t-1}) + \boldsymbol{H}_t(\boldsymbol{\Psi}_t - \boldsymbol{\Psi}_{t/t-1}), \qquad \boldsymbol{a}(\boldsymbol{\Psi}_t) \approx \boldsymbol{a}(\boldsymbol{\Psi}_{t/t}) + \boldsymbol{A}_t(\boldsymbol{\Psi}_t - \boldsymbol{\Psi}_{t/t}), \tag{3}$$

by denoting $\boldsymbol{H}_t = (\partial \boldsymbol{h}/\partial \boldsymbol{\Psi})|_{\boldsymbol{\Psi}_t = \boldsymbol{\Psi}_{t/t-1}}$ and $\boldsymbol{A}_t = (\partial \boldsymbol{a}/\partial \boldsymbol{\Psi})|_{\boldsymbol{\Psi}_t = \boldsymbol{\Psi}_{t/t}}$, the Jacobians of the measurement model and of the dynamic model. After some matrix manipulations, the a posteriori estimates and their error covariance

[2] OpenGL can perfectly deal with negative lights.

matrices $\boldsymbol{P}_{t/t}$ are given by the filtering equations:

$$\begin{aligned}\boldsymbol{K}_t &= \boldsymbol{P}_{t/t-1}\boldsymbol{H}_t^{\mathrm{T}}(\boldsymbol{R} + \boldsymbol{H}_t\boldsymbol{P}_{t/t-1}\boldsymbol{H}_t^{\mathrm{T}})^{-1},\\ \boldsymbol{\Psi}_{t/t} &= \boldsymbol{\Psi}_{t/t-1} + \boldsymbol{K}_t(\boldsymbol{s}_t - \boldsymbol{h}(\boldsymbol{\Psi}_{t/t-1})),\\ \boldsymbol{P}_{t/t} &= (I - \boldsymbol{K}_t\boldsymbol{H}_t)\boldsymbol{P}_{t/t-1},\end{aligned} \tag{4}$$

and the a priori estimates with their error covariance matrices $\boldsymbol{P}_{t+1/t}$ by the prediction equations:

$$\boldsymbol{\Psi}_{t+1/t} = \boldsymbol{a}(\boldsymbol{\Psi}_{t/t}), \qquad \boldsymbol{P}_{t+1/t} = \boldsymbol{A}_t\boldsymbol{P}_{t/t}\boldsymbol{A}_t^{\mathrm{T}} + \boldsymbol{Q}. \tag{5}$$

### 5.2. Interpretation of the equations

By considering the equations in (4), the filter produces $\boldsymbol{\Psi}_{t/t}$ by rectifying the predicted estimate $\boldsymbol{\Psi}_{t/t-1}$ by the correction term $\boldsymbol{K}_t(\boldsymbol{s}_t - \boldsymbol{h}(\boldsymbol{\Psi}_{t/t-1}))$, based on the difference between the real observations $\boldsymbol{s}_t$ and the predicted ones $\boldsymbol{h}(\boldsymbol{\Psi}_{t/t-1})$. An interesting interpretation of the Kalman filter behavior is that it iterates to make the system state vector $\boldsymbol{\Psi}_{t/t}$ *fit* both the observations $\boldsymbol{s}_t$ and the dynamic evolution model represented by the equations in (5).

This "differential fitting" interpretation helps to understand why, with a carefully chosen measurement model, the Kalman filter is able to align the synthesized images with the speaker's image even though the video camera is *not* calibrated: the idea is to derive the observation function from the synthesis operations, so that $\boldsymbol{s}_t = \boldsymbol{h}(\boldsymbol{\Psi}_t)$ is the vector of the 2D positions of the synthetic facial feature points in the image plane. In this case, $\boldsymbol{\Psi}_t$ includes the 6 degrees of freedom of the synthetic face in the synthetic world, and the filter will modify $\boldsymbol{\Psi}_t$ to output the 2D positions of the tracked facial features, thus aligning the speaker's face and the synthetic model.

#### 5.2.1. Dynamic model

The 6 parameters needed to render the synthetic head are

$$\mathscr{S} = (t_X, t_Y, t_Z, \alpha, \beta, \gamma)^{\mathrm{T}},$$

representing the 3 translations and the 3 rotations with respect to the $X$, $Y$ and $Z$ axes, and they are included with their first- and second-order time derivatives in the filter state vector,

$$\boldsymbol{\Psi} = (\mathscr{S}^{\mathrm{T}}, \dot{\mathscr{S}}^{\mathrm{T}}, \ddot{\mathscr{S}}^{\mathrm{T}})^{\mathrm{T}},$$

for the dynamic model of the system evolution, simply based on Newtonian physics under a constant acceleration assumption,

$$\mathscr{S}_{t+\mathrm{d}t} = \mathscr{S}_t + \dot{\mathscr{S}}_t\,\mathrm{d}t + \tfrac{1}{2}\ddot{\mathscr{S}}_t\,\mathrm{d}t^2. \tag{6}$$

#### 5.2.2. Measurement model

Although the viewing camera is not calibrated, the observation model of the Kalman filter mimics a perspective projection of focal length $\mathscr{F}$ (see Fig. 5). The actual value of $\mathscr{F}$ is not sensitive for the tracking algorithm, as the Kalman filter will adapt its state vector computation to fit the observations, as if they had been viewed by an ideal camera of this focal length.

Considering the rigid transformation applied by the OpenGL library on the 3D facial feature localization $(x, y, z)^{\mathrm{T}}$ in head mesh coordinates, the final 2D measurement $(x_{\mathrm{I}}, y_{\mathrm{I}})^{\mathrm{T}}$ in the image plane given the parameters

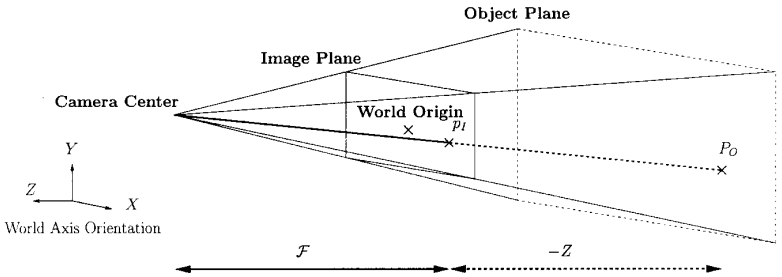


Fig. 5. Camera analogy for the view model, with $\boldsymbol{P}_O = (X, Y, Z, 1)^{\mathrm{T}}$, and $\boldsymbol{p}_\mathrm{I} = (x_\mathrm{I}, y_\mathrm{I})^{\mathrm{T}} = (X\mathcal{F}/(\mathcal{F} - Z), Y\mathcal{F}/(\mathcal{F} - Z))^{\mathrm{T}}$ (only the negative part of the $Z$ axis is seen by the camera).

$\mathcal{S} = (t_X, t_Y, t_Z, \alpha, \beta, \gamma)^{\mathrm{T}}$ and $(x, y, z)^{\mathrm{T}}$ is

$$\begin{pmatrix} x_\mathrm{I} \\ y_\mathrm{I} \end{pmatrix} = \begin{pmatrix} \dfrac{\mathcal{F}}{\mathcal{N}}(c_\beta c_\gamma x - c_\beta s_\gamma y + s_\beta z + t_X) \\ \dfrac{\mathcal{F}}{\mathcal{N}}((s_\alpha s_\beta c_\gamma + c_\alpha s_\gamma)x + (-s_\alpha s_\beta s_\gamma + c_\alpha c_\gamma)y - s_\alpha c_\beta z + t_Y) \end{pmatrix}, \tag{7}$$

with $\mathcal{N} = (-c_\alpha s_\beta c_\gamma + s_\alpha s_\gamma)x + (c_\alpha s_\beta s_\gamma + s_\alpha c_\gamma)y + c_\alpha c_\beta z + t_Z - \mathcal{F}$ (for clarity reasons, $c_\alpha$, $s_\alpha$, $c_\beta$, $s_\beta$, $c_\gamma$ and $s_\gamma$, respectively, stand for $\cos\alpha$, $\sin\alpha$, $\cos\beta$, $\sin\beta$, $\cos\gamma$ and $\sin\gamma$).

## 6. Tracking

### 6.1. Differential block-matching

The main complications encountered by a block-matching (or *pattern correlation*) algorithm are, on the one hand, the local geometric distortions, scales of facial features and changes of lighting due to the speaker's 3D motions, and on the other hand, the large rotations out of the image plane: in this case, the facial features might not fit in rectangular blocks, and the reference patterns are matched against portions of other objects in the background. Both issues are addressed by tracking synthetic facial features in the real speaker's view, as long as the block-matching procedure is able to allow some photometric differences between the synthetic and real patterns, and to be aware of the background pixels that are likely to interfere in the correlation score for extreme head poses. Extending the theory presented in [16], we propose that a classic differential block-matching formulation be adapted to be robust to the photometric model failures on synthesized features in Section 6.1.1, and support non-rectangular patterns and extreme orientations in Section 6.1.2.

#### 6.1.1. Photometric robustness

The differential block-matching algorithm is derived by considering that a reference pattern, denoted $\boldsymbol{I}(\boldsymbol{\mu} = \boldsymbol{0}, \tau = 0) = (p_1, \ldots, p_m)^{\mathrm{T}}$ (all pixels are placed in a column vector), undergoes some perturbations $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^{\mathrm{T}}$ (most often displacements in the image, with sub-pixel precision). Writing a Taylor series expansion for small perturbations between two consecutive frames, we have

$$\boldsymbol{I}(\boldsymbol{\mu}, \tau) \approx \boldsymbol{I}(\boldsymbol{0}, 0) + \boldsymbol{M}\boldsymbol{\mu} + \boldsymbol{I}_t\tau, \tag{8}$$

with

$$\boldsymbol{M} = \left( \frac{\partial \boldsymbol{I}}{\partial \mu_1}(\mathbf{0},0) \middle| \cdots \middle| \frac{\partial \boldsymbol{I}}{\partial \mu_n}(\mathbf{0},0) \right) \quad \text{and} \quad \boldsymbol{I}_t = \frac{\partial \boldsymbol{I}}{\partial t}(\mathbf{0},0).$$

Solving for $\boldsymbol{\mu}$ in the least mean square fashion yields

$$\boldsymbol{\mu} = -(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathrm{T}}\boldsymbol{I}_t, \tag{9}$$

which may be rewritten by identifying the lines of $-(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathrm{T}}$ as

$$\begin{pmatrix} \mu_1 \\ \vdots \\ \mu_n \end{pmatrix} = \left( \begin{array}{c} \boldsymbol{r}_1 \\ \hline \vdots \\ \hline \boldsymbol{r}_n \end{array} \right) \boldsymbol{I}_t. \tag{10}$$

In Eq. (8), the $\boldsymbol{\mu}$ perturbations are general enough to represent a local pattern rotation or scaling [11], and to integrate a luminance scale and offset perturbation in the case of a photometric model failure for a synthesized feature

$$\left( \frac{\partial \boldsymbol{I}}{\partial_{\text{lum. scale}}}(\mathbf{0},0) = \boldsymbol{I}(\mathbf{0},0) \quad \text{and} \quad \frac{\partial \boldsymbol{I}}{\partial_{\text{lum.offset}}}(\mathbf{0},0) = (1,\ldots,1)^{\mathrm{T}} \right).$$

The luminance offset degree of freedom defined in the photometric model (Eq. (1)) is taken into account, and need not be explicitly recovered during the matching procedure, since only the translation parameters $\mu_1 = \boldsymbol{r}_1\boldsymbol{I}_t$ and $\mu_2 = \boldsymbol{r}_2\boldsymbol{I}_t$ are computed from (9) to be given to the Kalman filter: the only overhead implied by additional degrees of freedom takes place during the computation of $-(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathrm{T}}$, but they do not mean more computations during the iterative estimation of $\mu_1$ and $\mu_2$.

*6.1.2. Background robustness*

Used together, the Kalman filter and the synthetic patterns can predict the background position near the participant's facial features due to the 3D pose: in Fig. 6, the synthesized model is rendered on a black background, which appears in the extracted rectangular nose pattern for this particular head orientation. The pattern pixels are classified into 2 subsets, $\boldsymbol{I}|_{\mathrm{F}}$ and $\boldsymbol{I}|_{\mathrm{B}}$, whether they belong to the face or to the background area.[3] We restrict the computation of the temporal gradient $\boldsymbol{I}_t$ in Eq. (10) to $\boldsymbol{I}|_{\mathrm{F}}$, so that the potential background objects in the real view have no impact on the correlation.

Hence, the synthetic patterns are turned into visual clues to match only the potential feature areas in the real speaker's image, and our tracking algorithm finds the correct best match despite other objects in the features neighborhood (see also Fig. 11 for extreme head rotations).

Restricting $\boldsymbol{I}_t$ in Eq. (10) to the $\boldsymbol{I}|_{\mathrm{F}}$ pattern subset, we get our modified block-matching algorithm giving the 2D displacement $(d_x, d_y)^{\mathrm{T}}$ of each feature area:

$$\begin{pmatrix} d_x \\ d_y \end{pmatrix} = \left( \begin{array}{c} \boldsymbol{r}_x \\ \hline \boldsymbol{r}_y \end{array} \right) \begin{pmatrix} g_1 \\ \vdots \\ g_m \end{pmatrix} \quad \text{with} \quad \begin{cases} g_i = p_i^{\text{real}} - p_i^{\text{synthetic}} & \text{if } p_i^{\text{synthetic}} \in \boldsymbol{I}|_{\mathrm{F}}, \\ g_i = 0 & \text{if } p_i^{\text{synthetic}} \in \boldsymbol{I}|_{\mathrm{B}}. \end{cases} \tag{11}$$

[3] In practice, the background is drawn in black at the encoder level within the feedback loop, to be easily segmented from the face regions.

Fig. 6. For this typical rotation, the synthetic background used by the encoder appears in the tracked nose area. As its pixels would lead to a wrong match in the real image, they are discarded by the tracker for the computation of the temporal gradient $\boldsymbol{I}_t$. As a result, the tracker restricts itself only to the potential face pixels in the video frame, and is able to find the right match despite the unknown background in the real world.

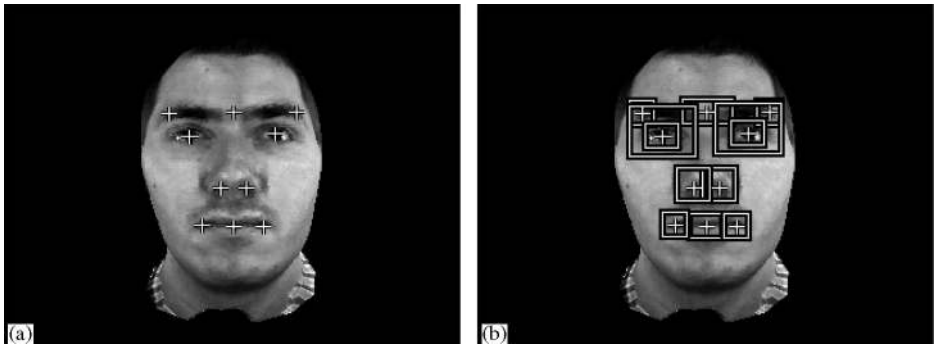


Fig. 7. The 12 facial features used to evaluate the tracking algorithm: (a) feature positions; (b) sizes of the tracked areas – Each eye represents 2 facial features, once with the eyebrows, and once without them.

## 7. Numerical evaluation of the global motion tracking

To assess the accuracy of the estimation of the global motion, we ran our algorithm on a synthetic video sequence, where the parameters to be recovered are known, using 12 tracked areas surrounding the main facial features, shown in Fig. 7. The face in this sequence oscillates around its initial position, undergoing respectively 3 translations along the $X$, $Y$ and $Z$ axes, 3 rotations around the same axes, and a combination of all degrees of freedom (see the video on the web [25]).

This sequence lasts 20 s, at 30 frames per second, for a total amount of 600 frames, and its resolution is $320 \times 242$ pixels. To make this evaluation more significant, a background image has been inserted behind the face model to mimick the real world case shown by Fig. 8. The images were rendered using an ideal perspective camera, with a focal length of 2. The global motion parameters used to synthesize the face model were obtained using a polynomial interpolator of third order between 23 keyframes, without enforcing the constant acceleration assumption of the Kalman filter. The maximum rotations are depicted in Fig. 9, and are large enough to verify the behavior of the tracking algorithm when the background is likely to interfere with the face patterns. Rotations are expressed in degrees, and translations in the face model coordinates: as the face vertices are scaled in the range $[-1, 1]$, a translation of $+3$ means a translation of about 1.5 times the face size.

Fig. 8. The real-world sequence, which inspired the synthetic sequence of Fig. 9 made to assess the tracking accuracy.

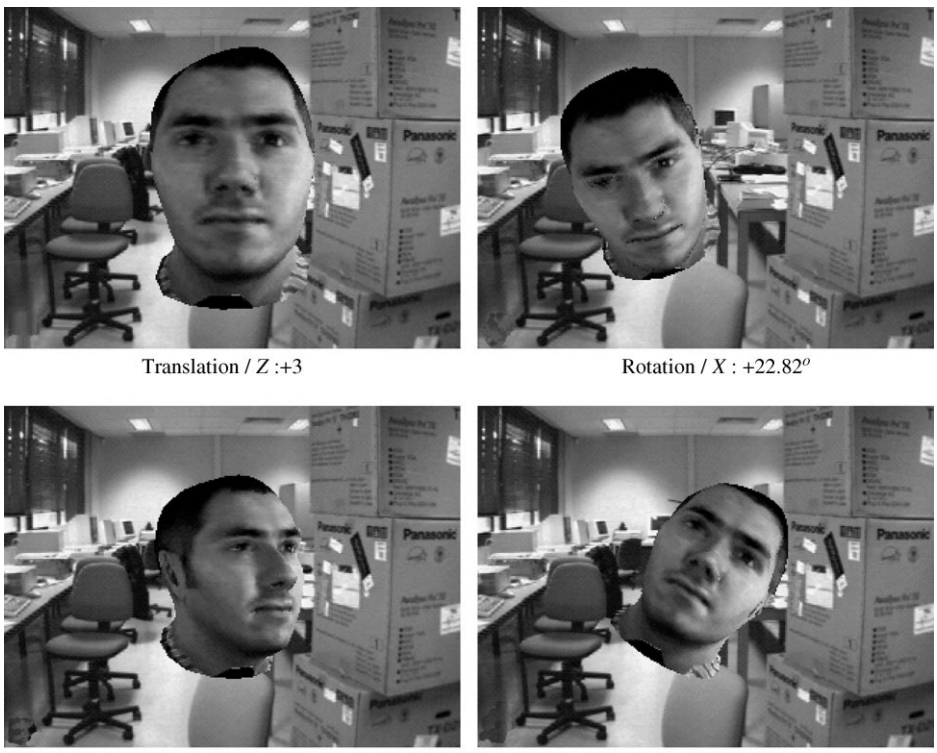


Fig. 9. Maximum amplitude of the parameters used to evaluate the tracking algorithm, with respect to the initial position. The $X$, $Y$ and $Z$ axes of rotation are local to the head model (i.e. they are not related to the camera).

We plotted the recovered parameters against the true ones for the 600 frames in Fig. 10. It can be seen that the algorithm successfully recovers the rigid motion of the face in the video sequence, but not surprisingly, the most difficult parameters to be recovered are the rotations around the $X$ and $Y$ axes, since they lead to non-affine diplacements in the 2D image plane, and are solved by the linearization of the observation model

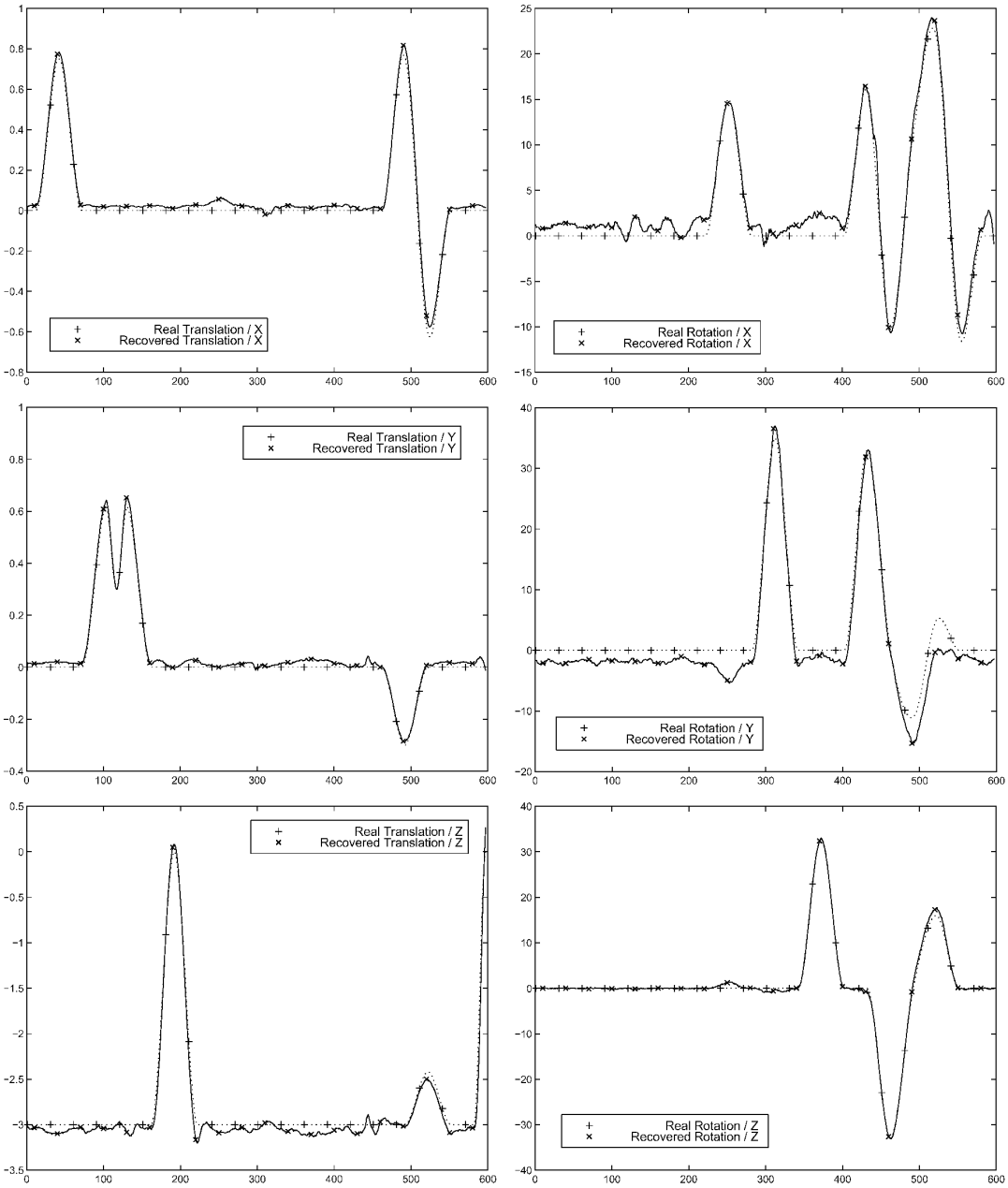


Fig. 10. Real and recovered $XYZ$ positions and orientations for the 600 frames.

computed by the Kalman filter at each frame. One may also notice that the recovered parameters are noisy, especially when the true parameters do not change. This is due to the Kalman filter, which does not sufficiently smooth out the a posteriori estimations. We could have obtained smoother results if we had incremented the noise level of the observations in the filter. However, there is always a compromise to be

Table 5
Statistics about the tracking accuracy on a synthetic test sequence

| | Max. Abs. Error | Mean error | Variance | $10\log_{10}$ SNR |
|---|---|---|---|---|
| Translation | | | | |
| $/X$ | 0.0681 (frame 510) | 0.0231 | $2.0372\times10^{-4}$ | 18.49 |
| $/Y$ | 0.0421 (frame 444) | 0.0147 | $8.1550\times10^{-5}$ | 19.70 |
| $/Z$ | 0.4235 (frame 589) | $-0.0567$ | $3.9956\times10^{-3}$ | 30.51 |
| Rotation | | | | |
| $/X$ | 3.4551 (frame 445) | 0.9130 | 0.5732 | 15.43 |
| $/Y$ | 5.4403 (frame 531) | $-1.9609$ | 1.5645 | 12.44 |
| $/Z$ | 1.7035 (frame 529) | 0.0606 | 0.2428 | 25.87 |

made between the smoothing capability of a Kalman filter (by decreasing the uncertainty of the predictions), and its reactivity to new situations (by decreasing the uncertainty of the incoming observations): for instance, with the noise settings used in this evaluation, it can be seen on the plots that the filter already tends to overestimate the parameters after a rapid transition (see frame 500 for the rotation around the $Y$ axis on Fig. 10) whereas the estimates lack some stability (from frame 0 to 300, when no rotation around the $Y$ axis should be detected).

Table 5 contains some statistics about the errors, and the signal to noise ratios (SNR) of the estimated parameters. They show that the global motion parameters are recovered quite well. The mean error is about 0.5% of the face scale for the translation parameters, and about 1 or 2 degrees for the rotation parameters, which makes the results quite fair. The largest errors are obtained during or after rapid transitions, indicating that the prediction model of the Kalman filter, assuming a constant acceleration, may not be the most appropriate.

## 8. Conclusions about the face-tracking algorithm

In the previous sections, we proposed a face-tracking framework which has the possibility to feed the feature-tracking procedure with synthesized facial patterns, implementing a *visual feedback loop*, thus solving the problems of lighting, scale and geometric deformations of the face implied by its rigid motion, with no explicit manipulation of the 3D geometry of the face model, unlike in other optical flow-based techniques. Our face cloning module is designed to work in real-world lighting, with few assumptions about the speaker's motion (allowing large rotations out of the image plane), and without marks or makeup on the user's face (see Fig. 11). To that purpose, we presented geometric and photometric modelling techniques to make the synthesized patterns closer to their real counterparts. We also reformulated a block-matching algorithm to make it work with synthetic input data, and showed how to handle the presence of the speaker's background in extreme positions. It is important to note that such an analysis/synthesis cooperation is successful because of the realism of our modelling techniques, and the design of the Kalman filter to drive the synthetic images. Our system is independent of any facial expression estimation algorithm, and can be used to generate a global animation stream for MPEG-4 encoders with life-like timings, leaving the task of facial expression animation to other methods.

The result of the tracking algorithm on a real face can be seen in a video sequence available on the WWW [23]. The original analyzed sequence lasted 30 s, and was captured in a $320\times242$ resolution at 10 frames per second. The analysis/synthesis speed mainly depends on the workstation graphics hardware acceleration and

Fig. 11. Head rotations supported by our face tracking system, needing neither markers nor uniform lighting.

its video acquisition speed. On a $O^2$ SGI workstation, the analysis frame rate using 12 facial feature areas is:

- 1 image per second, when synthesizing patterns, computing the product $-(\boldsymbol{M}^{\mathrm{T}}\boldsymbol{M})^{-1}\boldsymbol{M}^{\mathrm{T}}$, and updating the Kalman filter for every frame;
- 10 frames per second, when disabling synthetic pattern calculation for every frame, but still enabling the Kalman filter – in this case, large face rotations might cause the system to lose the user's head;
- full frame rate, when disabling both pattern synthesis and the Kalman filter – the system just tracks the facial features in 2D, without recovering the head 3D pose, and becomes very sensitive to rotations.

One of the questions that might be raised about the robustness of the tracking is what happens when the user closes his eyes, or has a facial expression that differs from the static expression of any feature area. In such a situation, some of the individual trackers may lose their facial feature, but there are enough tracked features to allow the outliers to be minimized by the Kalman filter. The prediction of the positions of the features in the next frame brings the trackers back in the right area until they produce correct matches. It is clear that the system will be even more robust if it can take into account the user's facial expressions, be they obtained by image analysis, or other means (like processings on the audio signal to infer the animation of the mouth).

### 8.1. Perspectives: synthesis of facial expressions

Now that we have a robust algorithm to track a human face and estimate its pose, our current work focuses on the synthesis of facial expressions (local animations), and their analysis by image processing only, still without any marker.

The face models constructed in Section 3 are unanimated. To efficiently generate facial expressions, we implemented several original or well-known animation techniques, operating on the 3D vertices, the texture coordinates attaching the vertices to the texture image, or the texture image itself:

*Mesh deformations*: by applying mesh morphing, it is possible to deform the face shape, to close the eyelids or squeeze the mouth of the face model (see Fig. 12(b));

*Displacements of texture coordinates*: extending the concept of mesh morphing to texture coordinates morphing, we can make the texture slide over the face model, without deforming the shape of the 3D wireframe. This technique is used to animate the eyebrows (Fig. 12(c));

*Texture sliding*: instead of adding new primitives for each eye-ball, we created holes in the texture image (via the transparency channel), and two separate textures behind the main one can be displaced to alter the model's gaze direction (Fig. 12(d));

*Texture blending*: beside moving some texture portions, it is possible to blend several textures together to produce a new one, for example to fade expression wrinkles into the model texture at low-cost in terms of real-time animation, instead of hard-coding them in heavy spline-based meshes (Fig. 12(e)).

All these techniques, of course, can be combined together to produce more complex facial expressions (see the video sequence [24]). We are now considering porting our displacements of texture coordinates and texture animations into an existing proprietary MPEG-4 viewer [36], in order to benefit from the streaming and networking capabilities of the standard (defined by the Digital Media Integration Framework, DMIF), while respecting the FAP syntax [29] to ensure our face models are MPEG-4 compliant.

### 8.2. Analysis of facial expressions

In conjunction, we are also investigating an original view-based approach to relate the analysis and the synthesis of facial expressions: using our realistic face model, we sample the visual space of facial expressions across various poses, via a set of $\boldsymbol{\mu}$ vectors containing facial animation parameters. Image patches for the facial features of interest (like the eyes, eyebrows and the mouth) are extracted, to produce distinct datasets of training examples. Then, a principal component analysis is performed over those training datasets, to extract a limited number of images optimally spanning the training space. These images (called *eigenfeatures*) allow us to characterize the facial expression of the user's face via a simple correlation mechanism, yielding a compact $\boldsymbol{\lambda}$ vector. And finally, a linear estimator will be designed to map the analysis scores $\boldsymbol{\lambda}$ to the face

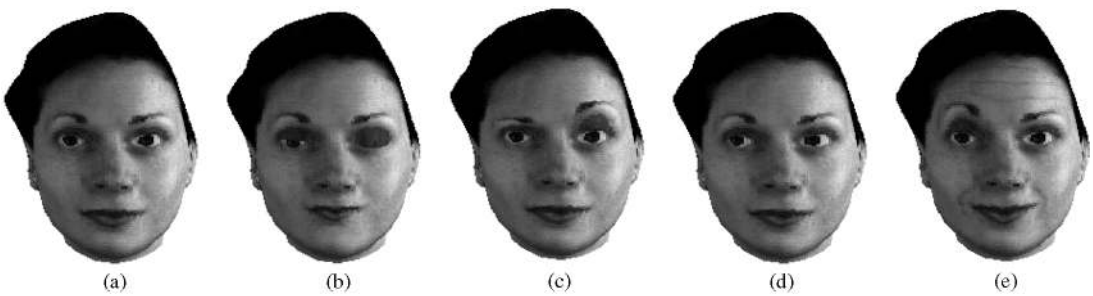


Fig. 12. Examples of facial animations: (a) neutral (initial) face model; (b) mesh deformations; (c) displacements of texture coordinates; (d) texture sliding; (e) texture blending.

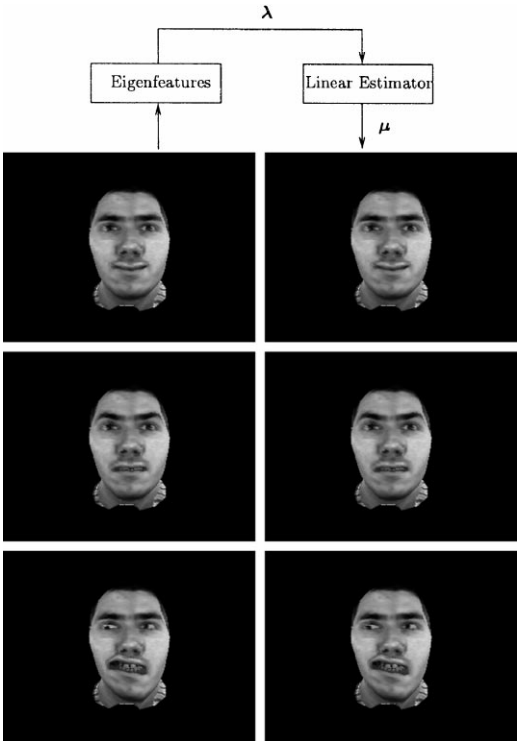


Fig. 13. Some analyses of facial expressions: each image of the left column was quantified by some eigenfeatures, giving a $\lambda$ vector. A linear estimator mapped $\lambda$ to the animation parameters $\mu$, which were rendered into the images of the right column.

animation parameters $\mu$. The major advantage of this approach is that all degrees of freedom permitted by the synthetic face can be precisely, automatically and systematically exploited by the training strategy.

Fig. 13 shows some preliminary results concerning the analysis of synthetic facial images using a linear estimator trained over the responses of the eigenfeatures, which are reported in [39]. We are currently extending this strategy to the analysis of real facial images.

## Acknowledgements

Eurécom's research is partially supported by its industrial members: Ascom, Cegetel, France Telecom, Hitachi, IBM France, Motorola, Swisscom, Texas Instruments, and Thomson CSF.

The authors would like to thank Hervé Delingette (INRIA Sophia-Antipolis) for his contribution in the model mesh construction; the University of Erlangen, Germany, and the L.U.A.P. (Université Paris-*VII*) for the original Cyberware scans; Danielle Pelé and Renaud Cazoulat (France Telecom R. & D.) for the MPEG-4 issue; and Katia Fintzel (Institut Eurécom) for her special appearance in Section 8.

## References

[1] P.M. Antoszczyszyn, J.M. Hannah, P.M. Grant, Accurate automatic frame fitting for semantic-based moving image coding using a facial code-book, in: IEEE International Conference on Image Processing, Lausanne, Switzerland, September 1996.

[2] P.M. Antoszczyszyn, J.M. Hannah, P.M. Grant, Automatic frame fitting for semantic-based moving image coding using a facial code-book, in: Eusipco-96, Trieste, Italy, September 1996.
[3] A. Azarbayejani, T. Starner, B. Horowitz, A. Pentland, Visually controlled graphics, IEEE Trans. Pattern Anal. Mach. Intell. 15 (6) (June 1993) 602–605.
[4] S. Basu, I. Essa, A. Pentland, Motion regularization for model-based head tracking, Technical Report 362, MIT Media Lab., 1996.
[5] P. Belhumeur, D. Kriegman, What is the set of images of an object under all possible lighting conditions?, in: IEEE Conference on Computer Vision and Pattern Recognition, November 1996.
[6] G. Bozdağı, M. Tekalp, L. Onural, 3-D motion estimation and wireframe adaptation including photometric effects for model-based coding of facial image sequences, IEEE Trans. Circuits Systems Video Technol. 4 (3) (June 1994) 246–256.
[7] CYBERWARE Home Page, URL http://www.cyberware.com
[8] D. DeCarlo, D. Metaxas, The integration of optical flow and deformable models with applications to human face shape and motion estimation, in: IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, California, 18–20 June 1996, pp. 231–238.
[9] H. Delingette, General object reconstruction based on simplex meshes, Int. J. Comput. Vision 32 (2) (1999) 111–146.
[10] J.-L. Dugelay, K. Fintzel, S. Valente, Synthetic natural hybrid video processings for virtual teleconferencing systems, in: Picture Coding Symposium, Portland, Oregon, 21–23 April 1999.
[11] J.-L. Dugelay, H. Sanson, Differential methods for the identification of 2D and 3D motion models in image sequences, Signal Processing: Image Communication 7 (1) (1995) 105–127.
[12] P. Eisert, B. Girod, Model-based 3D-motion estimation with illumination compensation, in: Sixth International Conference on Image Processing and its Applications (IPA 97), Dublin, Ireland, July 1997, pp. 194–198.
[13] P. Eisert, B. Girod, Analysing facial expressions for virtual conferencing, IEEE Comput. Graphics Appl. 18 (5) (September 1998) 70–78.
[14] A. Eleftheriadis, A. Jacquin, Automatic face location detection and tracking for model-assisted coding of video teleconferencing sequences at very low bit-rates, Signal Processing: Image Communication 7 (3) (June 1995) 231–248.
[15] K. Fintzel, J.-L. Dugelay, Visual spatialization of a meeting room from 2D uncalibrated views, in: H. Niemann, H.-P. Seidel, B. Girod (Eds.), Image and Multidimensional Digital Signal Processing'98, Alpbach, Austria, 1998, pp. 323–326.
[16] G. Hager, P. Belhumeur, Real-time tracking of image regions with changes in geometry and illumination, in: IEEE Conference on Computer Vision and Pattern Recognition, November 1996.
[17] Institut National de l'Audiovisuel, Televirtuality project: cloning and real-time animation system, URL http://www.ina.fr/INA/Recherche/TV
[18] T.S. Jebara, A. Pentland, Parametrized structure from motion for 3D adaptive feedback tracking of faces, in: IEEE Conference on Computer Vision and Pattern Recognition, November 1996.
[19] T. Kanade, P.J. Narayanan, P.W. Rander, Virtualized reality: concepts and early results, in: IEEE Workshop on Representation of Visual Scenes, Cambridge, MA, June 1995. In conjunction with ICCV'95.
[20] S.M. Kay, Fundamentals of Statistical Signal Processing Estimation Theory, PTR Prentice-Hall, Englewood Cliffs, NJ, 1993, Chapter 13, pp. 419–477.
[21] R. Koch, Dynamic 3-D scene analysis through synthesis feedback control, IEEE Trans. Pattern Anal. Mach. Intell. 15 (6) (June 1993) 556–568.
[22] H. Li, P. Roivainen, R. Forchheimer, 3-D motion estimation in model-based facial image coding, IEEE Trans. Pattern Anal. Mach. Intell. 15 (6) (June 1993) 545–555.
[23] MPEG demo of the face tracking system, URL http://www.eurecom.fr/ ~ image/TRAIVI/valente-8points.mpg (1,782,100 bytes).
[24] MPEG demo of the animation system, URL http://www.eurecom.fr/ ~ image/TRAIVI/animation.mpg (1,258,604 bytes).
[25] MPEG sequence used to evaluate the tracking algorithm, URL http://www.eurecom.fr/ ~ image/TRAIVI/accuracy-test.mpg (2,906,348 bytes).
[26] MPEG-4 synthetic/natural hybrid coding, URL http://www.es.com/mpeg4-snhc
[27] J. Neider, T. Davis, M. Woo, OpenGL Programming Guide, OpenGL Architecture Review Board, Release 1 Edition, June 1993.
[28] J. Ohya, Y. Kitamura, F. Kishino, N. Terashima, Virtual space teleconferencing: real-time reproduction of tridimensional human images, J. Visual Commun. Image Representation 6 (1) (March 1995) 1–25.
[29] Overview of the MPEG-4 standard, Technical Report ISO/IEC JTC1/SC29/WG11 N2725, International Organisation for Standardisation, Seoul, South Korea, March 1999.
[30] Proceedings of the Second International Conference on Audio- and Video-based Biometric Person Authentification (AVBPA'99), Washington, DC, 22–23 March 1999.
[31] M. Proesmans, L. Van Gool, One-shot 3D-shape and texture acquisition of facial data, in: Audio- and Video-based Biometric Person Authentification, Crans-Montana, Switzerland, March 1997, pp. 411–418.
[32] M.J.T. Reinders, P.L.J. van Beek, B. Sankur, J.C.A. van der Lubbe, Facial feature localization and adaptation of a generic face model for model-based coding, Signal Processing: Image Communication 7 (1) (1995) 57–74.
[33] A. Saulnier, M.-L. Viaud, D. Geldreich, Real-time facial analysis and synthesis chain, in: International Workshop on Automatic Face- and Gesture-Recognition, Zurich, Switzerland, 1995, pp. 86–91.

[34] A. Schödl, A. Haro, I.A. Essa, Head tracking using a textured polygonal model, in: Workshop on Perceptual User Interfaces, San Francisco, California, 4–6 November 1998.
[35] C. Schoeneman, J. Dorsey, B. Smits, J. Arvo, D. Greenberg, Painting with light, in: SIGGRAPH 93, Anaheim, California, 1–6 August 1993, pp. 143–146.
[36] J. Signès, J. Close, Streaming VRML and mixed media with MPEG-4, in: SIGGRAPH 98, Orlando, Florida, 19–24 July 1998.
[37] L. Tang, T.S. Huang, in: Automatic construction of 3D human face models based on 2D images, in: IEEE International Conference on Image Processing, Lausanne, Switzerland, September 1996.
[38] S. Valente, J.-L. Dugelay, A multi-site teleconferencing system using VR paradigms, in: Ecmast, Milano, Italy, 1997.
[39] S. Valente, J.-L. Dugelay, Face tracking and realistic animations for telecommunicant clones, in: IEEE International Conference on Multimedia Computing and Systems, Florence, Italy, 7–11 June 1999.
[40] VIDAS – VIDeo ASsisted audio coding and representation, URL http://miralabwww.unige.ch/Vidas-TO.html
[41] VRML, URL http://vrml.sgi.com