Performance-based Virtual Face Gestures

Ilse Ravyse and Hichem Sahli
Vrije Universiteit Brussel, Department ETRO
Audio-Visual Signal Processing Group
Interdisciplinary Institute for Broadband Technology
Pleinlaan 2, 1050 Brussel, Belgium
Email:{icravyse, hsahli}@etro.vub.ac.be

***Abstract*— The face motion plays a special role in sustaining the human gestures. The emerging media applications are conveying the human expressions by means of virtual faces which are typically controlled by a parameterized description. In this context, we developed a *Facial Analysis and Synthesis Scheme* able to extract a spatial-temporal semantic face description from an image sequence. This research proposes new methods for detection and tracking of the face and facial features, for estimating the face movements and the facial expressions, and for extracting efficiently stored facial animation. By combining vision, graphics and mechanical models, we can solve the inverse problem: estimating the 3-dimensional face deformation from the 2-dimensional image measurements. Animating the face model with the extracted parameters, results in a faithful reproduction of the original gestures.**

I. Introduction

Communication of the face motion is an important aspect of the human interaction. Its automatic extraction from an image sequence is a challenging task due to the complexity of the face gestures. It requires the measurement and modeling of all spatial-temporal semantical properties of the face in the scene. Different subproblems in a fully automatic vision-based face analysis system have to be distinguished [1]:

- Face localization, extraction or segmentation aims at finding the image position of a face. The goal of face detection is to determine whether or not any face is present in the image and, if so, to localize it. In this research we assume that a quasi-frontal facing person is present in the image.
- Face tracking methods continuously estimate the location of a face in an image sequence. Orientation, scale or other characteristics of the face target may be estimated as well. A prerequisite to tracking is the knowledge of the initial face location, which can be found by the face detection in the first image frame.
- The eyes, nose, nostrils, eyebrow, mouth, lips, ears, etc., are frequently called the 'face features'. It are the particular motions of all those features that make up a face gesture, no matter if this is a verbal (talking) or nonverbal (smile, surprise , ...) face motion. However, also the remaining face skin appearance contributes to these communicative actions and its motion must be grasped as well by the analysis system.
- Face recognition or face identification compares an input image against a face image database and reports a match. The purpose of face authentication is to verify the claimed identity of an individual in an image. Both problems will not be treated in this research as they do not relate to motion analysis of the face but rather to static face analysis. However, it must be said that they are part of the biometric research that is of increasing public interest due to the recent government security threats.
- Face expression or gesture recognition deals with the classification of the spatio-temporal information about the face and face features. Most of these systems use the human affective state classes (happiness, sadness, ...) [3], although the interest of behavioral psychological science and multimodal/media human computer interfaces prefer classifying local motions. This latter classes can be re-used (as model parameterizations) for the animation synthesis of a 3-dimensional face model as will be the subject of the following sections.

In this research, we propose the estimation of the face motion by the aid of a virtual face model that contains the related a-priori face knowledge. The sublime combination of computer vision, mechanical laws and computer graphics allows not only to analyze the rigid and nonrigid face motion, but at the same time it is capable to re-create the expressions on the 3-dimensional face model. Furthermore, the next generation video-compression standard MPEG4 [4], [5] intends to represent a face movie via a similar parameterized face model and foresees an efficient storage and transmission for it.

The techniques to extract the face gestures from a video recording of a talking actor and to recreate those gestures on a virtual face model are reviewed in Section II. First, the estimation of the different types of motion of the face are discussed in Sections II-A to II-D. Next, the synthesis of the extracted motion on the face model is highlighted in Section II-E. Concluding remarks and applications of the scheme are given in Section III .

II. Combined Analysis and Synthesis

We have set up a scheme that accomplishes the extraction of the parameters of the 3-dimensional face motion from a 2-dimensional recording of an actor's performance. The animation of a computer graphics face model is controlled

by two sets of parameters: 'static' parameters are foreseen for the face structure or geometry to construct the particular facial feature constellation; and 'dynamic' parameters are connected to the rigid motion and to the deformation of the face. This separation is reflected in the estimation of the facial expressions from the image sequence. The natural modeling of facial color, shape and apparent motion permits the retrieval of the valuable face gesture information lost in the 3- to 2-dimensional conversion inherent to the imaging process. This inverse problem has been solved at different levels, as explained in the following sections.

A. 2D Face Motion

Face extraction and head tracking exploit the facial color and spatial properties [6]. In the first image frame of a sequence, an ellipsoid head region is found by a personalized skin color detection and a shape-based face verification. Then, to follow this region in time, we have developed a tracking algorithm that models the joint probability of the location and the color of the detected face. The change of the face image location is modeled by an affine motion because this has the property to keep the ellipsoid shape after motion transformation and it can deal with scale changes which arise when the actor approaches the camera. The results of the estimation on the *Tom* sequence are shown by the ellipsoids in the Figure 1. This new method achieves a better robustness than the color-histogram trackers.

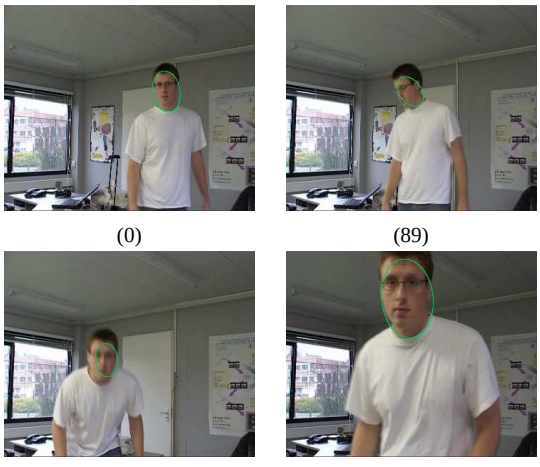

Fig. 1. The *Tom* image sequence: ellipsoid face detection at frame 0, and tracking results of frames 89, 156 and 250

B. Facial Feature Gestures

Automatic gesture analysis is provided by the image filtering of a mathematical morphology scale-space [7]. An eye blink is observed as a change in the amount and size of the gray-scale details in a column of a local image window, as displayed in the eye drawings of Figure 2. The face motion related to this gesture can thus be represented as a change of the facial feature's shape.

Fig. 2. Details in the shape of an opened and closed eye

This information about the eye (or mouth) opening is contained in the local morphological *sieved* scale images. By measuring the difference between those scale images and normalizing between the observations during the entire sequence, a feature shape parameter is assembled. In this way, we have realized a direct extraction of the spatio-temporal gesture parameters that control the virtual face's eye and mouth opening. It is a fast technique that can be used to observe the evolution of the facial features. Figure 3 displays an input sequence and a face model with the same eye and mouth opening. In the middle of this Figure, the local images around the eyes and mouth are shown with a bar-representation of the extracted parameters.

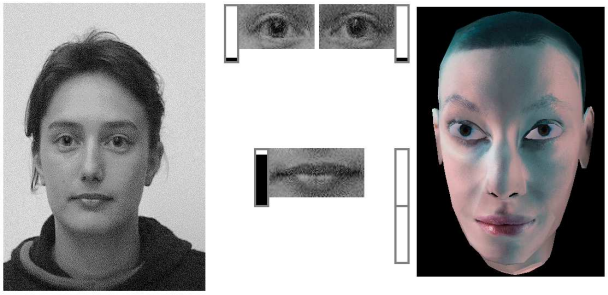

Fig. 3. Extraction result of eye and mouth gestures: (a) input image; (b) image windows with a bar that indicates if the feature is opened (empty bar) or closed (black bar), the double bar gives the orientation of the face (based upon the image eye locations); (c) the virtual face model

C. Face Model Representation

In this research, a virtual face model that approximates the surface structure of a human face is not only used in synthesis, but as well in the analysis of the semantical face gestures. This requires a 3D wire-frame or mesh of which the topology follows the anatomical characteristics of a generic face (not only in curvature, but also in the density and ordering of the constituting triangles in highly deformable face regions). Scene calibration prepares such a generic face model geometry to support the motion. Using the semantic correspondences between the face image and the model, a camera calibration positions the model in the 3-dimensional face scene, and a structure calibration [8] makes the model adaptations to person's looks. The selection of the face morphing algorithm to alter the face model's vertex positions which does not distort the face model has been based on a new sensitivity study of the radial basis function interpolation. Calibration of a face model on the *Claire* sequence is shown in Figure 4.

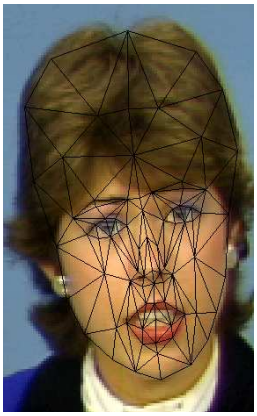

Fig. 4. calibrated face model

D. 3D Face Motion

The estimation of the 3-dimensional facial expressions is an ill-posed problem due to the high degree of deformation of the face surface [9]. It cannot be solved without imposing constraints on the face motion. In this work, a new estimation approach [10] integrates a natural motion model using the muscle forces caused by elastic skin displacements (implemented using a finite element model) [11]. We took into account the motion projection by linking the optical flow measurement with the modelled scene flow in a nonlinear least-squares formulation, and by giving a physical motivation to the regularization. Smooth deformations of the face model have been obtained, as depicted in Figure 5.

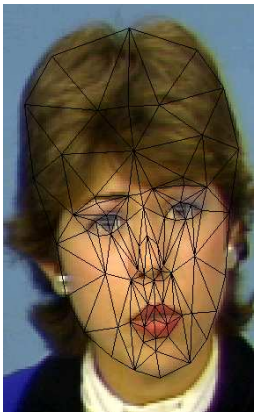

Fig. 5. nonrigid motion applied on the face model, viewed on the image showing the face gesture.

E. Face Motion Parameters

The illusion of motion of a 3D virtual object is created by changing the shape of its visible surface. This process is governed by the deformation of its model geometry. Animation is done by specifying the motion controls to deal with input animation parameter data that describes the intended (extracted) deformation. Additionally, texture mapping can enhance the face's realistic appearance. In this research, we have focussed on how the automatic controls for the face surface deformation are established and represented.

The techniques for face animation roughly belong to 3 groups: model fitting, muscle action modeling, and physically-based modeling. All techniques have in common that the animation becomes feasible trough grouping vertices into face regions with controlled motion paths. Since a 3D model of an object quickly contains more than a few tens of vertices, the overview of their displacement would easily be lost otherwise.

1) **Model fitting**:
The goal is to generate new face expressions by combining a limited set of known face looks. Face target shapes or parameters can be used to alter the generic face shape. A library of shapes with different expressions are created from a single model [12]. Interpolation methods are applied to go from one such geometry state of the 3D model to another (without changes in its topology). Because the face gestures are merely the geometrical resemblances of the true gestures, the designer of the face model has a big responsibility in creating the key face shapes and the smooth transitions between them.

2) **Muscle action modeling**:
For human body shapes, articulated *skin & bones* models use stick segments connected by rotating joints and covered with a blob of stretching skin surface [13]. The skeleton is moved with inverse kinematics, available in commercial animation packages, and a skin layer slides over the body bones. The animator specifies in the end position of the joints, and an algorithm computes the motion path of the segments. Unfortunately, the anatomy of a face does not consist of such segments, so this technique cannot be adopted. An approximate model of the physical/mechanical processes involved in the face deformation has been introduced by Waters and Terzopoulos [14]. They defined the face surface deformations with a local interpolation function for animating a muscle contraction. The usage of dynamic parameters that semantically correspond to face muscles is still commonly used for many virtual face models. The drawback is that the muscle structure must be provided by the artist that created the face model.

3) **Physically-based modeling**:
The face behavior can be simulated with mechanical laws to provide the parametrization. This does not only involve the model surface, but also the underlying interaction mechanics of skeleton, muscles and tissue. All forces that are acting on the face are taken into account. A deeper insight in the mechanics of the face is not required for the model designer, because the gestures are the result of the simulation of a physical process of external and internal influences on the face material. In combination with the mechanical parameter extraction, as discussed in section II-D, this opens a large study area. This type of mechanical modeling, applied on

elastic objects, is currently a hot topic in the research on graphical simulation environments.

Key-framing is the most widely spread animation technique and can be applied directly on the model geometry as in list item 1. It is based on the (experimental) knowledge that the human visual system is able to form a coherent view of dynamic scenes from a rapid sequence of still images. In key-framing the animators make sequential drawings of a continuing action and project its photographs at a constant frame rate onto a screen. The idea is that the animator only specifies the pose of the character at a certain point of time, and again for another point in time some number of frames later. These key-frames or key-poses represent the most important characteristic captured poses in the motion to draw, as shown in Figure 6. In-betweens are all other captured poses that are required to make smooth motion and to fill in the yet undrawn frames (at a certain frame rate). The key-framing technique is a very labor-intensive approach for animation because complete specification of the character is required at each key-frame.

For computer animation, other techniques have been developed: control-based animation uses a parametrization of the character. Currently all commercial facial animation schemes can be viewed as such parameterizations. Therefore the graphically represented face animations benefit from control rules which can be easily applied to a face model [15]. We have realized this by the dynamic application of the MPEG4 face animation parameters (FAP) with the radial basis function interpolation [4], [5] and by the FAP calculation from the previously extracted deformations via a least-squares fit.

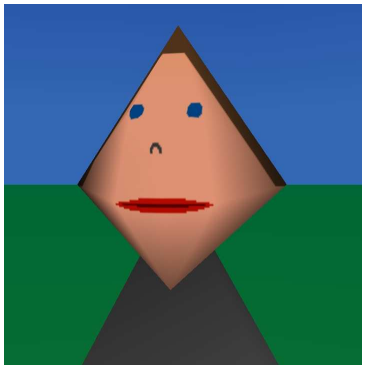

Fig. 6. A key frame of a virtual face model

III. Conclusion

In this paper we have investigated the extraction of face motion parameters from an image sequence in an extendible, well-founded framework. The 3-dimensional gestures of the face of a person can be recovered and re-played by a virtual face model.

Gesture recognition to support human-computer interfaces is currently investigated in computer vision, computer graphics and sociological research groups worldwide. Especially the mechanical and physically-based models, as investigated in this paper, bring a valuable asset in natural interaction with face puppets or animations by means of combined arts, computer vision and computer graphics. As continuation of this research work an educational learning environment for pronunciation training of a set of spoken words will be investigated. The visual mouth shape recognition and synthesis will be combined with visual speech unit, called viseme, recognition from an acoustic signal for multi-modal face analysis.

Applications in the emerging media which convey this synthetic content via a virtual human are numerous, amongst them are the interactive television, games, tele-presence, news-reading on internet and intuitive user-interfaces.

Acknowledgment

This research was partially sponsored by the *Instituut voor aanmoediging van Innovatie door Wetenschap en Technologie in Vlaanderen* (IWT) as a specialization grant.

References

[1] M.-H. Yang, D. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 34–58, Jan. 2002.

[2] W. Y. Zhoa, R. Chellappa, A. Rosenfeld, and P. Phillips, "Face recognition: A literature survey," *ACM Computing Surveys (CSUR)*, vol. 35, pp. 399 – 458, Dec. 2003.

[3] M. Pantic and L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1424–1445, Dec. 2000.

[4] ISO/IEC, "Jtc 1/sc 29/wg 11 n2501 coding of moving pictures and audio, information technology - generic coding of audio-visual objects, part1:systems, final draft international standard." Atlantic City, Oct. 1998.

[5] ISO/IEC, "Jtc 1/sc 29/wg 11 n2502 coding of moving pictures and audio, information technology - generic coding of audio-visual objects, part2:visual, final draft international standard." Atlantic City, Oct. 1998.

[6] I. Ravyse, V. Enescu, and H. Sahli, "Kernel-based head tracker for videophony," in *The IEEE International Conference on Image Processing 2005 (ICIP2005), Genoa, Italy, 11-14/09/2005*, vol. 3, pp. 1068–1071, Sep. 2005.

[7] I. Ravyse, H. Sahli, M. Reinders, and J. Cornelis, "Eye activity detection and recognition using morphological scale-space decomposition," in *15th International Conference on Pattern Recognition, ICPR2000 (September 3-8, 2000) Barcelona, Spain*, vol. 1, pp. 1080–1083, Sep. 2000.

[8] F. Lavagetto, R. Pockaj, and M. Costa, "Smooth surface interpolation and texture adaptation for mpeg-4 compliant calibration of 3d head models," *Image and Vision Computing Journal*, vol. 18, no. 4, pp. 345–354, 2000.

[9] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 137–154, Mar. 2005.

[10] I. Ravyse, *Facial Analysis and Synthesis*. PhD thesis, Vrije Universiteit Brussel, Dept. Electronics and Informatics, Belgium, May 2006.

[11] Y. Bascar and D. Weichert, *Nonlinear continuum mechanics of solids: fundamental mathematical and physical concepts*. 193 p., Springer-Verlag: Berlin, 2000. ISBN 3-540-66601-X.

[12] E. Chuang, *Analysis, Synthesis, and Retargeting of Facial Expressions*. PhD thesis, Stanford University, Mar. 2004.

[13] I. Salomie, A. Gavrilescu, I. Ravyse, R. Deklerck, J. Cornelis, and M. Preda, "Efficient animation of meshgrid models, using skin & bone system," in *3rd IEEE Benelux Signal Processing Symposium SPS2002 (21-22/03/2002), Leuven, Belgium*, pp. 21–24, Mar. 2002.

[14] K. Waters, "A muscle model for animating three-dimensional facial expression," *Computer Graphics ACM*, vol. 21, pp. 17–24, Jul. 1987.

[15] F. I. Parke and K. Waters, *Computer Facial Animation*. A K Peters, 1996. ISBN 1-56881-014-8.