Making Machines Understand Facial Motion & Expressions Like Humans Do

Ana C. Andrés del Valle & Jean-Luc Dugelay

Multimedia Communications Dpt. Institut Eurécom
2229 route des Crêtes. BP 193. Sophia Antipolis. France
{andres, dugelay}@eurecom.fr

Abstract

Complete interaction amongst humans and machines unavoidably needs computers to understand human emotions. Most emotive information comes from facial motion and expression. This article presents the design of a new procedure for image analysis that is able to understand facial actions on monocular video sequences, without imposing restrictions on the speaker or its environment. The exposed technique follows a global-to-specific analysis approach that tries to imitate the way people analyze face motion: by dividing this analysis in processes of different level of detail.

1 Introduction

Researchers from the Computer Vision, Computer Graphics and Image Processing communities have been studying the difficulties associated with the analysis and synthesis of faces under motion for more than 20 years. The analysis and synthesis techniques being developed can be useful for the definition of low-rate bit image compression algorithms (model-based coding), new cinema technologies as well as for the deployment of virtual reality applications, videoconferencing, to enforce with realism the users' presence in games, etc. As computers evolve towards becoming more human oriented machines, human-computer interfaces, behavior-learning robots and disable adapted computer environments will use face expression analysis to be able to react to human action.

Facial analysis of motion and expression from monocular (single) images is widely investigated and of main interest because non-stereoscopic static images and videos are the most affordable and extensively used in visual media. In this field, research oriented towards the improvement of communications for the hearing impaired, like the video phone application presented by Sarris and Strintzis (2001), prove that understanding facial motion is of great help. We must not forget that "(…) many standard facial expressions, such as shaking the head for 'no' or raising the eyebrows to form a question, are used extensively to convey emotion, emphasis and intensity." (Disability Online, 2001).

Until recently, the analysis of rigid and non-rigid facial motion has been studied separately. Solutions given to deploy general expression analysis, mainly based on only-image processing techniques such as template matching or image Principal/Independent Component Analysis, are often designed to work on faces viewed from a frontal perspective and under controlled environment conditions. Tian, Kanade and Cohn (2000), already present a way to adapt their near-to-front analysis scheme to other head poses. They define "multiple state models", where different facial component models are used for different head poses; this solution proves to be toilsome. Approaches that search for more detailed facial action information need to be well aware of the

pose of the head not only in reference to the image plane but also to its physical location. The 3D pose tracking of the head allows better control of the analysis and more freedom of movement to the speaker; we can appreciate these advantages in those approaches that fit a 3D mesh to better track face expressions (Ogata, Murai, Nakamura, & Morishima, 2001). It also permits more efficient image information retrieval by allowing, for instance, the rectification of the analyzed image to a known frontal pose (Chang, Chen, C.-C., Chou, & Chen, Y.-C., 2000). Nevertheless, most solutions end up having their own limitations related to the conditions under which the face is being analyzed. Ideally, proper facial motion analysis should work under any circumstances, regardless of the characteristics of the face and its surroundings. The design of a robust and potentially improvable analysis scheme can be obtained by observing human behaviour at performing this same task.

People can understand facial action even when faces are under very bad lighting or in the presence of disturbing objects over them. This is basically due to the fact that humans are able to automatically reduce the complexity of the analysis into different parts and to do this analysis progressively. First, our sight adapts to the lighting conditions under which the face is and we decide if further understanding is possible; then, we locate the head and get its rigid motion (its pose) and finally, we pay attention to the different details of the face that are interesting to us because they contain expression information. When humans are not able to perform an exhaustive analysis (lighting is very bad, or a significant part of the face is occluded), they make up for the missing information (generally assuming standard human behavior) or they simply accept that they cannot understand the face motion they are observing. In this article we present a complete system that performs facial motion and expression analysis on monocular video sequences trying to simulate this natural and intuitive human conduct. The methods and algorithms presented are designed to work under natural and non-controlled situations: not assuming the use of a calibrated camera, with no need of precise lighting conditions or markers on the person; to be as universal as possible, trying to avoid any system training specific to an individual previous to the analysis; to be as precise as the analysis conditions permit, allowing the user as much freedom of movement as possible and by generating an analysis stratagem permitting potential improvement in its precision; and to potentially work in real-time, thus permitting instant expression understanding from the interpretation of coherent facial animation parameters.

2 Overall Analysis Scheme

We consider face analysis from a video sequence as a function of the general pose of the face on the sequence, the illumination conditions under which the video is recorded and the face expression movements. We believe that synthesizing the analyzed face action on the speaker's clone[1] is a relevant way to check that the extracted motion information is correct. Following such premise, our analysis scheme attempts to obtain facial animation parameters (FAP) allowing as much motion precision as possible during their later synthesis, given the current conditions of analysis.

To obtain animation parameters from video frames in a robust manner, we need to be aware of the lighting on the face as well as of the physical nature of the features we are analyzing; this information will enable our algorithms to work under any lighting conditions and to remain robust throughout the analysis. We first estimate the pose of the face obtaining translation and rotation parameters of the head. And then, we extract some specific features from the face (eyes, eyebrows

1.1

[1] Clone: 3D synthetic representation of a person that not only represents realistically its appearance but also has the potential of being animated replicating exactly this person's facial motion

and mouth) and we apply on them some dedicated analysis techniques to obtain face animation parameters.
We utilize a two step process to develop our image analysis algorithms. First, we design image processing techniques to study the features extracted from a frontal point of view of the face. Faces show most of their expression information under this pose and this allows us to verify the correct performance of the image processing involved. Second, we extend our algorithms to analyze features taken at any given pose. This adaptation is possible because the motion models utilized during the analysis can be redefined in 3D space and the accuracy of the retrieved pose parameters is such that enables us to reinterpret the data we obtain from the image analysis in 3D.
Controlling the pose permits understanding the evolution of the 2D regions we are analyzing on the image and thus allows us to foresee if doing the analysis over a specific feature will be profitable. It also let us the possibility of designing hierarchical analysis algorithms with different levels of detail depending on the visibility and the quality of the feature. Combining motion information coming from the analysis of different features to deduce the most suitable feature action is one of the best ways to avoid incoherent and unnatural analysis results. The analysis final check should be done in layers and compensate for that information that may be missing from the analysis (e.g. from occluded parts). In our case, we start by making sure that both eyes are behaving in the same way, then, we will introduce the eyebrow motion and finally we will analyze the overall expression when including the mouth. Although this kind of approach limits expression behavior to standard, natural human face motion, it becomes very helpful in situations where face analysis may be difficult.

3 Head-Pose Tracking

To fully understand global head actions in front of a camera, we needed to develop a head tracking algorithm capable of determining accurately the pose in a 3D reference system. Many tracking algorithms tend to work using local 2D image reference systems on which they can only estimate the user's position on the screen (e.g. Bradski's Cam Shift algorithm (1998)). For the sequence of processes to follow: detection of the features to be analyzed, analysis and interpretation of the analyzed results, this information is not accurate enough.
To obtain precise information about the person's location in space, we have developed an algorithm that utilizes a feedback loop inside a Kalman filter. Kalman filtering has been applied to head tracking giving very positive results (Cordea, Petriu, E. M., Georganas, D., Petriu, D. C., & Whalen, T. E., 2001; Ström, 2002) and it enables the prediction of the translation and rotation parameters of the head from the 2D tracking of specific points of the face on the image plane.
The natural drawback of this system is the need of 3D information about the shape of the head we are tracking, that is, we must use a model that provides the 3D coordinates of the points whose projection is tracked on the image and fed to the filter to obtain the prediction of the pose parameters. Very often, a general head model is used; although this apparent drawback can be become a strong advantage if a realistic 3D synthetic representation of the user is available. In (Valente, & Dugelay, 2001), we showed that improvement in the amount of freedom of movement in front of the camera is possible if using the speaker's clone during the tracking. In our approach, the algorithm operating on the image plane extracts the 2D features to be tracked on the video sequence from the synthesized image of the model, onto which the predicted pose parameters have already been applied thus providing an adjusted view of the user in its future pose. Since this approach compares head models and video frames at the image level, models have to be an accurate 3D representation of the speakers, in shape and texture. Furthermore, some lighting compensation must be done on the synthetic world to adapt it to the illumination conditions of the video sequence. Details in how this is performed can be found on the aforementioned reference.

4 Facial Feature Analysis Development for a Frontal Point of View

Designing specific image processing algorithms for each of the facial features being analyzed enables us adapting the design following anatomical and muscular constraints. These constraints take into account the influence of lighting on the appearance of the feature and the restrictions natural human facial movements impose. We have designed motion models that control feature behavior in a frontal position. These models can easily be adapted to scale their complexity and amount of retrieved data depending on the analysis conditions (i.e. size of image, lighting, etc.)
This section briefly reviews the main characteristics of the algorithms.

- Eye-state analysis (Andrés del Valle & Dugelay, 2001):

The image processing analysis strategy decomposes the eye tracking actions in two categories: the open-close movement and the eyeball movement. To best exploit the physical characteristics of the eyes, a different algorithm analyzes each action. We define the degree of eye opening as proportional to the inverse of the amount of skin contained within the analyzed feature image. Gaze orientation is related to the position of the pupil on the feature region. Pupils can be defined as the lowest energy points of the eyes. We set a tight cooperation between the two analysis techniques in a temporal state analysis, which allows us to correct possible erroneous results from the algorithms.

- Eyebrow motion analysis (Andrés del Valle & Dugelay, 2002):

To study eyebrow behaviour from video sequences we utilize a new image analysis technique based on an anatomical-mathematical motion model. This technique conceives the eyebrow as a single curved object (arch) that is subject to the deformation due to muscular interactions. The action model defines the simplified 2D (vertical and horizontal) displacements of the arch. Our video analysis algorithm recovers the needed data from the arch representation to deduce the strength of the parameters involved.

- Mouth motion analysis (Andrés del Valle, Perales & Dugelay, 2003):

We have studied and observed the muscular and bone interaction during mouth motion, so not only lips are tracked but also tongue and teeth are taken into account. We have mathematically modeled mouth muscular interaction to deduce which are the minimum needed control points that will permit to synthetically replicate mouth motion through the understanding of its behavior.

5 Coupling Feature Analysis and Pose Tracking

The solution we propose (Andrés del Valle & Dugelay, 2002) defines the feature regions to be analyzed and the parameters of a motion analysis on 3D, over a head model in its frontal position. The complete procedure goes as follows (see Figure 1):
(i) We define and shape the area to be analyzed on the video frame. To do so, we project the 3D-regions of interest (ROIs) defined over the head model on the video image by using the pose parameters obtained from the rigid motion tracking, thus getting the 2D-ROIs.
(ii) We apply each one of the previous image processing techniques on these areas extracting the data required for the motion models.
(iii) We interpret these data from a three-dimensional perspective by inverting the projection and the transformations due to the pose (data pass from 2D to 3D) assuming all data fall on the same image plane. At this point, we can compare the results with the feature analysis parameters already predefined on the neutral head model and decide which has been the feature action.

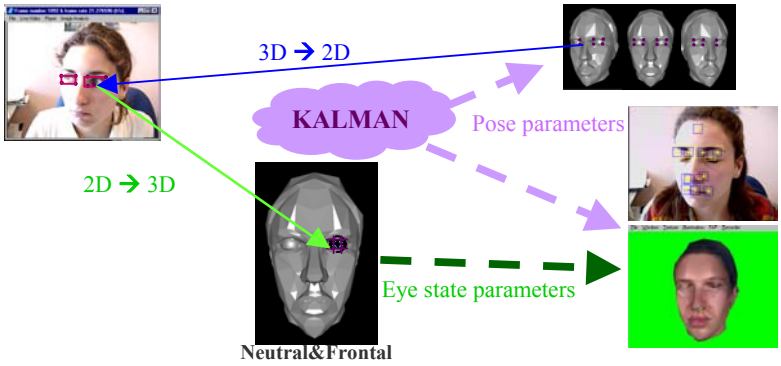

Figure 1: This diagram depicts the different interactions amongst 2D and 3D data during the analysis

6 Conclusions

In this article, we have presented an original approach to tackle the challenging issue of analyzing facial behavior on monocular video input in environments with minimum constraints (without known lighting or specific head pose). In the proposed framework, the complexity of the overall head action understanding is split into the study of several subproblems. Pose and face features are first analyzed independently and then jointly. This allows us to validate the robustness and performance of the algorithms involved because we are able to clearly detect and identify the origin and the nature of unexpected inaccurate results. Therefore it also permits to control the improvement of our algorithm steps while we are developing the different modules.

References

Andrés del Valle, A. C., Perales, F. J. & Dugelay, J.-L. (2003). *Analysis of mouth and lip motion and its coupling with pose tracking*. Technical Report to be published jointly by Eurecom Institute and the Universitat de les Illes Balears.

Andrés del Valle, A. C. & Dugelay, J.-L. (2002) Eyebrow movement analysis over real-time video sequences for synthetic representation. *Lecture Notes in Computer Science*. Springer (Ed.), Vol. 2492, 213–225.

Andrés del Valle, A. C., & Dugelay, J.-L. (2002). Facial expression analysis robust to 3D head pose motion. *Proceedings of the IEEE International Conference on Multimedia and Expo.*

Andrés del Valle, A. C. & Dugelay, J.-L. (2001). Eye state tracking for face cloning. *Proceedings of the IEEE International Conference on Image Processing.*

Bradski, G.R.(1998). Computer vision face tracking as a component of a perceptual user interface. *Workshop on Applications of Computer Vision*, 214–219.

Chang, Y.-J., Chen, C.-C., Chou, J.-C., & Chen, Y.-C. (2000) Virtual Talk: a model-based virtual phone using a layered audio-visual integration. *Proceedings of the IEEE International Conference on Multimedia and Expo*.

Cordea, M. D., Petriu, E. M., Georganas, N. D., Petriu, D. C., & Whalen, T. E. (2001). 3D head pose recovery for interactive virtual reality avatars. *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*.

Disability Online. (2001). Auslan is a sign language. Retrieved January 21, 2003, from http://www.disability.vic.gov.au/dsonline/dsarticles.nsf/pages/Auslan_is_a_sign_language?OpenDocument

Ogata, S., Murai, K., Nakamura, S., & Morishima, S. (2001) Model-based lip synchronization with automatically translated synthetic voice toward a multi-modal translation system. *Proceedings of the IEEE International Conference on Multimedia and Expo.*

Ström, J. (2002, October). Model-based real-time head tracking. *Eurasip Journal on Applied Signal Processing*, 2002 (10), 1039–1052.

Tian, Y., Kanade, T., & Cohn, J. F. (2001, February). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2), 97–115.

Valente, S., & Dugelay, J.-L. (2001). A visual analysis/synthesis feedback loop for accurate face tracking. *Signal Processing: Image Communication*, 16(6), 585–608.