Biological Motion of Speech

Gregor A. Kalberer[1], Pascal Müller[1], and Luc Van Gool[1,2]

[1] D-ITET/BIWI, ETH Zurich, Switzerland,
[2] ESAT/PSI/Visics, KULeuven, Belgium
kalberer,mueller,vangool@vision.ee.ethz.ch

Abstract. The paper discusses the detailed analysis of visual speech. As with other forms of biological motion, humans are known to be very sensitive to the realism in the ways the lips move. In order to determine the elements that come to play in the perceptual analysis of visual speech, it is important to have control over the data. The paper discusses the capture of detailed 3D deformations of faces when talking. The data are detailed in both a temporal and spatial sense. The 3D positions of thousands of points on the face are determined at the temporal resolution of video. Such data have been decomposed into their basic modes, using ICA. It is noteworthy that this yielded better results than a mere PCA analysis, which results in modes that individually represent facial changes that anatomically inconsistent. The ICs better capture the underlying, anatomical changes that the face undergoes. Different visemes are all based on the underlying, joint action of the facial muscles. The IC modes do not reflect single muscles, but nevertheless decompose the speech related deformations into anatomically convincing modes, coined 'pseudo-muscles'.

Introduction

Humans are all experts at judging the realism of facial animations. We easily spot inconsistencies between aural and visual speech, for instance. So far, it has been very difficult to perform detailed, psychophysical experiments on visual speech, because it has been difficult to generate groundtruth data that can also be systematically manipulated in three dimensions. Just as is the case with body motion, discrete points can give useful information on speech [12]. Nevertheless, the authors of that study concluded that '... point-light stimuli were never as effective as the analogous fully-illuminated moving face stimuli'.
In general two classes of 3D facial analysis and animation can be destinguished - physically based (PB) and terminal analog (TA) [8]. In contrast to the PB class, that involves the use of physical models of the structure and function of the human face, the TA class cares only about the net effect (a face surface) without resorting to physically based constructs.
In this paper, we follow the TA strategy, because this strategy has the advantage that correspondences between different faces and different movements stand by at any time. Furthermore, for animation the mere outer changes in the polygonal face shapes can be carried out faster than muscle and tissue simulations. As

human perception also only takes visible parts of the face into account, such a simplification seems justified.
We propose a system that measures in 3D the detailed facial deformations during speech. The data are quite detailed in that thousands of points are measured, at the temporal resolution of video. Several contributions in this direction have already been undertaken, but with a substantially smaller number of points (see e.g. Pighin *et al.* [10], Reveret *et al.* [11], Lin *et al.* [7] and Guenter *et al.* [1]). But as the aforementioned psychophysical experiments have demonstrated, it is important to have control over more detailed data when setting up experiments about the visual perception of speech.
The paper also analyses the extracted 3D dynamics. The data are decomposed into basic deformation modes. Principal Component Analysis yields modes that are anatomically inconsistent, but Independent Components are better able to split the deformations up into modes that make sense in their own right. This suggests that they are also better able to home in on the kind of actions facial muscles excerce on the face. They could be considered to each represent a 'pseudo-muscle', of which the actions can be linearly combined to yield realistic speech. Such animations have actually been tried, with good results. The animation aspects have been discussed elsewhere [4–6].

1 Extracting 3D face deformations

This section describes how groundtruth data were acquired by observing real, talking faces. People where asked to read sentences, with a sufficient variety of phonemes.
For the 3D shape extraction of the talking face, we have used a 3D acquisition system that uses structured light [3]. It projects a grid onto the face, and extracts the 3D shape and texture from a single image. By using a video camera, a quick succession of 3D snapshots can be gathered. The acquisition system yields the 3D coordinates of several thousand points for every frame. The output is a triangulated, textured surface. The problem is that the 3D points correspond to projected grid intersections, not corresponding, physical points of the face. Hence, the points for which 3D coordinates are given change from frame to frame. The next steps have to solve for the physical correspondences.

1.1 Mapping the raw data onto a face topology

Our approach assumes a specific topology for the face mesh. This is a triangulated surface with 2268 vertices for the skin, supplemented with seperate meshes for the eyes, teeth, and tongue (another 8848, mainly for the teeth).
The first step in this fitting procedure deforms the generic head by a simple rotation, translation, and anisotropic scaling operation, to crudely align it with the neutral shape of the example face. In order to correct for individual physiognomies, a piecewise constant vertical stretch is applied. This transformation

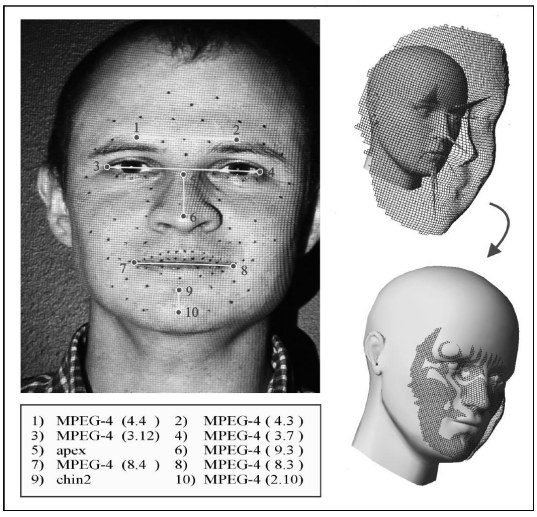

Fig. 1. A first step in the deformation of the generic head to make it fit a captured 3D face, is to globally align the two. This is done using 10 feature points indicated in dark grey in the left part of the figure. The right part shows the effect: patch and head model are brought into coarse correspondence.

minimizes the average distance between a number of special points on the example face and the model (10 points; they are indicated in black in figure 1). These have been indicated manually on the example faces, but could be extracted automatically [9]. A final adaptation of the model consists of the separation of the upper and lower lip, in order to allow the mouth to open. This first step fixes the overall shape of the head and is carried out only once (for the neutral example face). The result of such process is shown in the right column of figure 1.

The second step starts with the transformed model of the first step and performs a local morphing. This morphing maps the topology of the generic head model precisely onto the given shape. This process starts from the correspondences for a few salient points. This set includes the 10 points of the previous step, but also 106 additional points, all indicated in black in figure 2. Typically, the initial frame of the video sequence corresponds to the neutral expression. This makes a manual drag and drop operation for the 116 points rather easy. At that point all 116 points are in good correspondence. Further snapshots of the example face are no longer handled manually. From the initial frame the points are tracked automatically throughout the video, and only a limited manual interaction was necessary.

The 3D positions of the 116 points served as anchor points, to map all vertices of the generic model to the data. The result is a model with the shape and expression of the example face and with 2268 vertices at their correct positions. This mapping was achieved with the help of Radial Basis Functions.

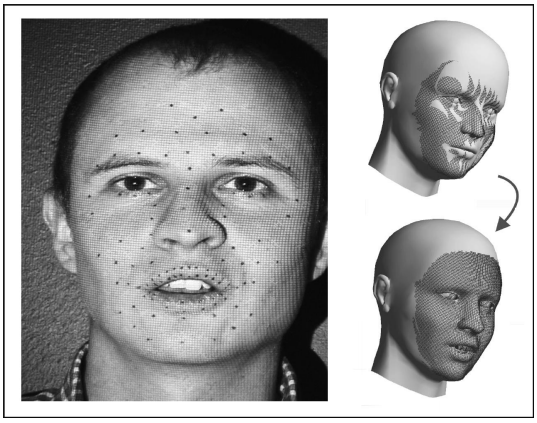

Fig. 2. To make the generic head model fit the captured face data precisely, a morphing step is applied using the 116 anchor points (black dots) and the corresponding Radial Basis Functions for guiding the remainder of the vertices. The right part of the figure shows a result.

Radial Basis Functions (RBFs) have become quite popular for face model fitting [10, 9]. They offer an effective method to interpolate between a network of known correspondences. RBFs describe the influence that each of the 116 known (anchor) correspondences have on the nearby points in between in this interpolation process. Consider the following equations

$$\boldsymbol{y_{i_{new}}} = \boldsymbol{y_i} + \sum_{j=1}^{n} w_j \boldsymbol{d_j} \tag{1}$$

which specify how the positions $\boldsymbol{y_i}$ of the intermediate points are changed into $\boldsymbol{y_{i_{new}}}$ under the influence of the n vertices $\boldsymbol{m_j}$ of the known network (the 116 vertices in our case). The shift is determined by the weights w_j and the virtual displacements $\boldsymbol{d_j}$ that are attributed to the vertices of the known network of correspondences. More about these displacements is to follow. The weights depend on the distance of the intermediate point to the known vertices:

$$w_j = h(s_j/r) \qquad s_j = \|\boldsymbol{y_i} - \boldsymbol{m_j}\| \tag{2}$$

for $s_j \leq r$, where r is a cut-off value for the distance beyond which h is put to zero, and where in the interval $[0, r]$ the function $h(x)$ is of one of two types:

$$h_1 = 1 - x^{log(b)/log(0.5)} \qquad b \approx 5 \tag{3}$$

$$h_2 = 2x^3 - 3x^2 + 1 \tag{4}$$

The exponential type is used at vertices with high curvature, limiting the spatial extent of their influence, whereas the hermite type is used for vertices in a region

of low surface curvature, where the influence of the vertex should reach out quite far. The size of the region of influence is also determined by the scale r. Three such scales were used (for both RBF types). These scales and their spatial distribution over the face vary with the scale of the local facial structures.
A third step in the processing projects the interpolated points onto the extracted 3D surface. This is achieved via a cylindrical mapping. This mapping is not carried out for a small subset of points which lay in a cavity, however. The reason is that the acquisition system does not always produce good data in these cavities. The position of these points should be determined fully by the deformed head model, and not get degraded under the influence of the acquired data.
The interior of the mouth is part of the model, which e.g. contains the skin connecting the teeth and the interior parts of the lips. Typically, scarcely any 3D data will be captured for this region, and those that are captured tend to be of low quality. The upper row of teeth are fixed rigidly to the model and have already received their position through the first step (the global transformation of the model, possibly with a further adjustement by the user). The lower teeth follow the jaw motion, which is determined as a rotation about the midpoint between the points where the jaw is attached to the skull and a translation. The motion itself is quantified by observing the motion of a point on the chin, standardised as MPEG-4 point 2.10.
It has to be mentioned at this point that all the settings like type and size of RBF's, as well as whether vertices have to be cylindrically mapped or not, are defined only once in the generic model as attributes of its vertices.

1.2 Decomposing the data into their modes

Fig. 3. Four of the sixteen Principal Components, in order of descending importance (eigenvalue).

Principal Component Analysis probably is the most popular tool to analyse the relevant variability in data. A PCA analysis on the observed, 3D deformations has shown that 16 components cover 98.5% of the variation, which seems to suffice. When looking at the different Principal Components, several of them could not represent actual face deformations. Such components need to be combined

with others to yield possible deformations. Unfortunately, this is difficult to illustrate with static images (Fig.3), as one would have to observe the relative motions of points. Indeed, one of the typical problems was that areas of the face would be stretched in all directions simultaneously to an extent never observed in real faces (e.g. in the cheek area).

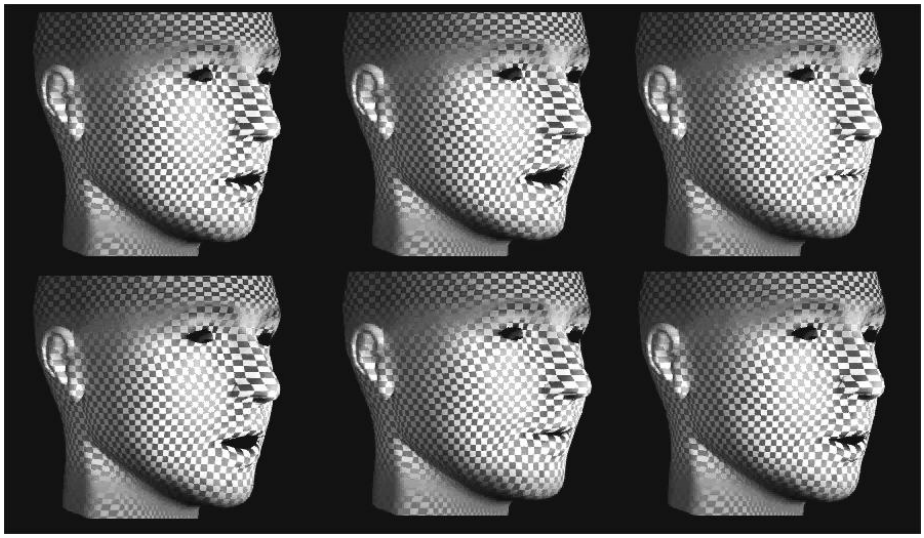

Fig. 4. Six of the sixteen Independent Components.

Independent Component Analysis (our implementation of ICA follows that propounded by Hyvärinen [2]), on the other hand, has yielded a set of modes that are each realistic in their own right. In fact, PCA is part of the ICA algorithm, and determines the degrees of freedom to be kept, in this case 16. ICA will look for modes (directions) in this PC space that correspond to linear combinations of the PCs that are maximally independent, and not only in the sense of being uncorrelated. ICA yields directions with minimal mutual information. This is mathematically related to finding combinations with distributions that are maximally non-Gaussian: as the central limit theorem makes clear, distributions of composed signals will tend to be more Gaussian than those of the underlying, original signals. The distributions of the extracted independent components came out to be quite non-Gaussian, which could clearly be observed from their χ^2 plots. This observation corroborated the usefulness of the ICA analysis from a mathematical point of view.

A face contains many muscles, and several will be active together to produce the different deformations. In as far as their joint effect can be modeled as a linear combination of their individual effects, ICA is *the* way to decouple the net effect again. Of course, this model is a bit naive, but nevertheless one would

hope that ICA is able to yield a reasonable decomposition of face deformations into components that themselves are more strongly correlated with the facial anatomy than the principal components. This hope has proved not to be in vane. Fig.4 shows 6 of the 16 independent components. Each of the Independent Components would at least correspond to a facial deformation that is plausible, whereas this was not the case for the Principal Components.

Finally, on a more informal score, we found that only about one or two PCs could be easily described, e.g. 'opening the mouth'. In the case of ICs, 6 or so components could be described in simple terms. When it comes to a simple action like rounding the mouth, there was a single IC that corresponds to this effect, but in the case of PCs, this rounding is never found in isolation, but is combined with the opening of the mouth or other effects. Similar observations can be made for the other ICs and PCs.

2 Conclusions

In this paper, we have described an approach to extract groundtruth data of the biological motion corresponding to 3D facial dynamics of speech. Such data are a prerequisite for the detailed study of visual speech and its visemes. The paper also discussed the variability found in the deformation data, and it was argued that ICA seems to yield more natural and intuitive results than the more usual PCA.

2.1 Acknowledgments

This research has been supported by the ETH Research Council and the EC IST project MESH (www.meshproject.com) with the assistance of our partners Univ. Freiburg, DURAN, EPFL, EYETRONICS, and Univ. of Geneva.

References

1. Guenter B., Grimm C., Wood D., Malvar H. and Pighin F., "Making Faces", *SIGGRAPH'98 Conf. Proc.*,vol. 32,pp. 55-66, 1998.
2. Hyvärinen A., "Independent Component Analysis by minimizing of mutual information", *Technical Report A46, Helsinki University of Technology*,1997.
3. http://www.eyetronics.com
4. Kalberer G. and Van Gool L., "Lip animation based on observed 3D speech dynamics", *SPIE Proc.*,vol. 4309,pp. 16-25,2001.
5. Kalberer G. and Van Gool L., "Face Animation Based on Observed 3D Speech Dynamics" *Computer Animation 2001. Proc.*,pp. 20-27,2001.
6. Kshirsagar S., Molet T. and Magnenat-Thalmann N., "Principal components of expressive speech animation", *Computer Graphics Int. Proc.*,pp. 38-44,2001.
7. Lin I., Yeh J. and Ouhyoung M., "Realistic 3D Facial Animation Parameters from Mirror-reflected Multi-view Video", *Computer Animation 2001 Conf. Proc.*,pp. 2-11,2001.
8. Massaro D. W., "Perceiving Talking Faces", *MIT. Press*,1998.

9. Noh J. and Neumann U., "Expression Cloning", *SIGGRAPH'01 Conf. Proc.*,pp. 277-288,2001.
10. Pighin F., Hecker J., Lischinski D., Szeliski R. and Salesin D., "Synthesizing Realistic Facial Expressions from Photographs", *SIGGRAPH'98 Conf. Proc.*,pp. 75-84,1998.
11. Reveret L., Bailly G. and Badin P., "MOTHER, A new generation of talking heads providing a flexible articulatory control for videorealistic speech animation", *ICSL'00 Proc.*,2000.
12. Rosenblum, L.D. and Saldaña, H.M, "Time-varying information for visual speech perception", In *Hearing by Eye*,vol. 2,pp. 61-81,ed. Campbell R., Dodd B. and Burnham D.,1998.