Face Analysis for the Synthesis of Photo-Realistic Talking Heads

Hans Peter Graf[1], Eric Cosatto[1], Tony Ezzat[2]
[1]AT&T Labs-Research, 100 Schulz Drive, Red Bank, NJ 07701, USA {hpg,eric}@research.att.com
[2]MIT, 45 Carleton Street E25-204, Cambridge, MA 02141, tonebone@ai.mit.edu

Abstract

This paper describes techniques for extracting bitmaps of facial parts from videos of a talking person. The goal is to synthesize photo-realistic talking heads of high quality that show picture-perfect appearance and realistic head movements with good lip-sound synchronization. For the synthesis of a talking head, bitmaps of facial parts are combined to form whole heads and then sequences of such images are integrated with audio from a text-to-speech synthesizer. For a seamless integration of facial parts into an animation, their shape and visual appearance must be known with high accuracy. The recognition system has to find not only the location of facial features, but must also be able to determine the head's orientation and recognize the facial expressions.

Our face recognition proceeds in multiple steps, each with an increased precision. Using motion, color and shape information, the head's position and the location of the main facial features are determined first. Then smaller areas are searched with matched filters, in order to identify specific facial features with high precision. From this information a head's 3D orientation is calculated. Facial parts are cut from the image and, using the head's orientation, are warped into bitmaps with 'normalized' orientation and scale.

1. Introduction

Animated characters, and talking heads in particular, are playing an increasingly important role in computer interfaces. An animated talking head attracts immediately the attention of a viewer, it can make a task more engaging and adds entertainment value to an application. Generating animated talking heads that look like real people is a very challenging task, and so far all synthesized heads are still far from reaching this goal. To be considered natural, a face has to be not only photo-realistic in appearance, but must also exhibit realistic head movements, emotional expressions, and proper plastic deformations of the lips, synchronized with the speech. We are trained since birth to recognize faces and facial expressions and, therefore, are highly sensitive to the slightest imperfections in a talking face.

Many approaches exist for modeling the human head [1] achieving different degrees of photo-realism and flexibility. Recently there has been a surge of interest in sample-based techniques (also referred to as data-driven modeling) for synthesizing photo-realistic heads. These techniques combine real images or parts of videos to generate new, animated sequences. For that purpose a person is recorded while speaking and then the whole face or parts of it are extracted from the video. A new sequence of a talking head is synthesized by integrating such parts into new faces.

The main difficulty for this approach is that large numbers of images need to be analyzed to generate a set of data samples that allow synthesizing lively sequences. A very high precision of the recognition is required to ensure that synthesized faces look natural and that the lip and head movements are smooth. When integrating facial parts into a new face, they have to be placed with an accuracy of typically less than one pixel; otherwise artifacts will be visible. Since recorded parts all vary in orientation and scale, their appearance has to be compensated for these effects. Hence, not only the precise shape, location and appearance of a facial part, but also the orientation of the head has to be measured.

Other researchers have used multiple cameras and markers on the face to find facial features and derive the 3D geometry of the head [2][3]. These systems have shown some impressive animations of facial expressions, yet they have not been demonstrated for speech production. A talking-head synthesis technique based on recorded mouth sequences of tri-phones (3 subsequent phonemes) has been demonstrated by Bregler et al. [4]. It parameterizes the samples only with the acoustic information, which limits the appearances in new sequences that can be synthesized. Ezzat et al. [5] have demonstrated a sample-based talking head system that uses morphing to generate intermediate appearances of mouth shapes from a very small set of manually selected mouth samples. This system produces very smooth transitions between mouth shapes, yet does not take coarticulation into account. Cosatto et al. [6] presented a sample-based talking head that uses several layers of 2D bit-planes as a model. Neither facial parts nor the whole head are modeled in 3D and therefore the system is limited in what new expressions and movements it can synthesize.

In our approach, we attempt to reduce some of the

limitations of previous systems by combining 3D modeling of the head with sample-based techniques. The system allows synthesizing speech, and provides a limited range of head movements and emotional expressions. Coarticulation is derived from recorded samples, resulting in naturally looking lip movements. Special emphasis has been put on keeping the recording process for the data samples easy and on generating them with minimal human intervention. We use only a single video camera and all features are extracted automatically without having to place any markers in the face. We also let the recorded person move the head. This makes the extraction of facial parts more challenging, but provides an environment where natural articulation and head movements, which typically accompany speech, can be recorded.

2. Model

A key problem with sample-based techniques is to control the number of image samples that need to be recorded and stored. A face's appearance changes due to talking, emotional expressions and head orientation, leading to a combinatorial explosion in the number of different appearances. To keep the number of samples at a manageable level we divide the face into a hierarchy of parts and model each part independently. Our face model is defined as follows:

2.1. Hierarchy of parts

The head is separated into a 'base face' and a number of facial parts. The base face covers the area of the whole face serving as a substrate onto which the facial parts are integrated. The facial parts are: mouth with cheeks, jaw, eyes, and forehead with eyebrows (Figure 1). Nose and ears are not modeled separately, but are part of the base face.

2.2. 3D model

The shape of each facial part is approximated with a small number of planes. This set of planes is used as a guide to map the facial parts onto the base face in a given pose (Figure 1). The positions and orientations of these planes follow the movements of the head, yet their shapes remain constant, even when the corresponding facial parts undergo non-rigid deformations. Hence, a model plane is more like a local window onto which a facial part is projected than a polygon of a traditional 3D model.

2.3. Sample bitmaps

For each facial part, sample bitmaps are recorded that cover the range of possible appearances produced by non-rigid deformations. For the base face, bitmap samples are recorded with the head in different orientations. The range of head rotations we consider at the moment is ±10° from a frontal view.

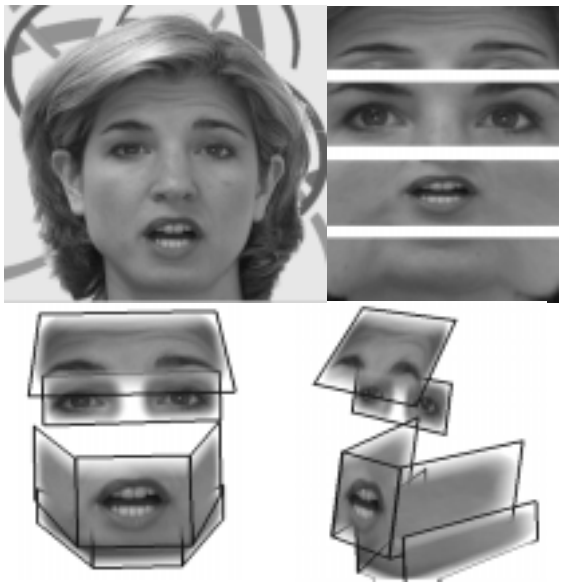

Figure 1. *Top left: The recorded face; top right: Normalized facial parts; bottom left: The 3D head model with bitmaps; bottom right: The 3D head model strongly rotated to illustrate the 3D shape of the model*

To generate a face with a certain mouth shape and emotional expression, the proper bitmaps are chosen for each of the facial parts. The head orientation is known from the base face, so that we can adapt the bitmaps onto the base face by warping. This operation is similar to traditional texture mapping. The difference is that for non-rigid deformations we select different bitmaps, rather than trying to squeeze one single bitmap into a new shape.

2.4. Capturing sample bitmaps

A head model is instantiated in two steps. First a few measurements are made on the subject's face to determine its geometry. For this we measure the positions of eye corners, nostrils, mouth corners and the bottom of the chin. Using these measurements, the model planes are adapted for each facial part.

In the second step for each facial part bitmaps are extracted from videos. A person is recorded while speaking freely, and the text is chosen to cover the most frequent triphones of the English language. The lip shapes depend not only on the phoneme articulated at any moment, but also on the context before and after. This phenomenon is known as coarticulation and has to be taken into account in order to synthesize naturally looking speech.

Once recorded, each frame of the videos is analyzed to

determine the head location and its 3D orientation. Then the facial parts are cut out and are normalized, i.e. they are warped into a new shape to compensate for different head orientations and scales. This normalization is necessary in order to characterize and compare the appearance of facial parts from different sequences.

Each facial part is labeled for an easy identification. For example, a mouth is labeled with geometric parameters, such as its width and the position of upper and lower lip. In addition we store the phonetic context where it was recorded plus a few parameters describing its appearance.

Because we are interested in capturing the natural behavior of the speaker, which includes typical head movements during speech and some emotional expressions, we try to keep the capture process as simple and non-intrusive as possible. Thanks to robust recognition algorithms we do not need any special markers on the subject's face, but rather exploit the natural richness in features of the face.

Figure2: *Image analysis for locating the head and the main facial features. Top row: Shape analysis with bandpass filtered image (left) and thresholded image (center). Bottom row: Optical flow (left), thresholded flow field (center); the image on the right shows the head outline (from color segmentation), the main facial features and the area of largest motion (chin area).*

3. Recognition

Sample-based synthesis of talking heads depends on a reliable and accurate recognition of the face and the positions of the facial features. The main challenge for the face recognition system is the high precision with which the facial features have to be located. An error as small as a single pixel in the position of a feature distorts the pose estimation of the head noticeably. To achieve such a high precision we analyze an image in three steps, each with an increased accuracy. Moreover, a large number of features are measured, providing an over-determined system of equations for the computation of the head pose so that errors in the individual features can be averaged.

The first step finds a coarse outline of the head plus estimates of the positions of the major facial features. In the second step the areas around the mouth, the nostrils and the eyes are analyzed in more detail. The third step, finally, zooms in on specific areas of facial features, such as the corners of the eyes, of the mouth and of the eyebrows and measures their positions with high accuracy. The recognition works with four distinct types of algorithms, namely: color analysis, motion analysis, texture/shape analysis and matched filters.

3.1. Locating the face

In a first step the whole image is searched for the presence of heads, and their locations are determined.

Color analysis: The first type of analysis is a color segmentation to find the areas with skin colors and colors representative of the hair. We use the hue for this segmentation, since the hue tends to be fairly constant across the face, regardless how the shadows fall. Only when the camera goes into saturation, there is a significant change in the hue of the skin. The training of the color segmentation is done by a leader clustering, using a set of 30 images where the segmentation has been done manually. The color analysis outputs binary blobs, identifying the areas of skin and hair.

Shape/Texture analysis: The second type of analysis segments the image based on textures and shapes. This analysis uses only the luminance of the image. First, the image is filtered with a band-pass filter, removing the highest and lowest spatial frequencies. The filters are adjusted to pass spatial frequencies typical for the mouth and eye areas of a face. Then a morphological operation followed by adaptive thresholding results in a binary image. The filter parameters are learned by analyzing a training set of labeled images. The analysis produces, like the color analysis, binary blobs marking areas of facial features (Figure 2).

Motion analysis: The motion within the picture is estimated with an optical flow algorithm. We use either an algorithm based on a block matching or, alternatively, an algorithm similar to the one described by Lucas and Kanade [7]. By thresholding the output of the optical flow, we identify areas where large displacements occur and mark them with binary blobs (Figure 2).

The color analysis as well as the texture and motion analysis produce blobs of connected binary pixels, marking areas where a facial feature may be present. This representation is a compact way of describing shapes and their relative positions. Features, such as width, height, position, and moments are measured from these blobs and are evaluated with classifiers. For example, an area marked by the color analysis as a candidate of a face is

combined with candidates of eye areas produced by the texture analysis. If relative sizes and positions match closely those of a reference face, this combination is evaluated further and combined with other features. Otherwise it is discarded.

3.2. Locating facial features

Finding the exact dimensions of the facial features is more challenging, since the person being recorded is moving the head and is changing facial expressions while speaking. This can lead to great variations in the appearance of a facial feature and can also affect the lighting conditions. For example, during a nod a shadow may fall over the eyes. Therefore, the analysis described above does not always produce accurate results for all facial features and we need to analyze further the areas around eyes, mouth, and the lower end of the nose.

The algorithm proceeds by analyzing the color space, periodically retraining it with a small number of frames. For example, the area around the mouth is cut out from five frames and, using a leader-clustering algorithm, the most prominent colors in the area are identified. By analyzing the shapes of the color segments we can assign the colors to different parts, such as the skin, the mouth cavity, the teeth and the lips (Figure 3). By repeating the color calibration periodically, we keep track of changes in the appearances of the facial features.

The shape analysis is also adapted to the particular facial feature under investigation by adjusting the filter parameters to the size and shape of a feature. In this way, the combination of texture and color analysis produces reliable measurements of the positions and outlines of the facial features.

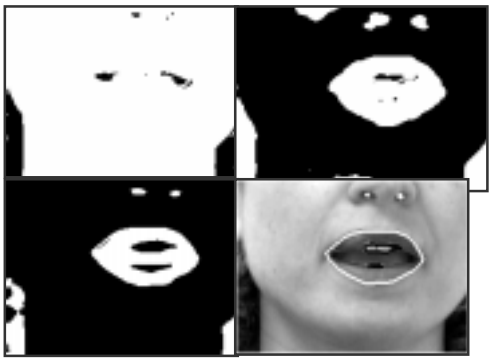

Figure 3*: Color segmentation of the mouth.*

Errors made by the system are of two types. The first type is a complete failure to identify a facial feature and the second type is inaccuracy in the measurements. A failure to identify a feature happens in about 1% of the frames, mostly when the head moves over a wider range than what was seen in the training images or when hair covers part of an eye. The accuracy achieved for the dimensions of the mouth are typically ±2 pixels (standard deviation), where the width of the mouth is around 100 pixels.

3.3. High accuracy feature points

For calculating the 3D head pose, a few points in the face have to be measured with high accuracy, preferably with an error of less than one pixel. The techniques described above tend to produce variations of, for example, ±2 pixels for the eye corners. Filtering over time can improve these errors significantly, yet a more precise measurement is still preferable.

We therefore add a third level of analysis to measure a few feature points with the highest accuracy. From a training set of 300 frames a few representative examples of one feature point are selected. For example, for measuring the position of the left lip corner, nine examples are selected (Figure 4). These samples are chosen based on the dimensions of the mouth. This means that the training procedure selects mouth images with three different widths and three different heights. From those images the areas around the left corner are cut out. For analyzing a new image, one of these sample images is chosen, namely the one where the mouth width and height are most similar, and this kernel is scanned over an area around the left half of the mouth.

To measure the similarity between the kernel and the area being analyzed, both are filtered with a high-pass filter before multiplying them pixel by pixel or calculating the pixel difference. This correlation identifies very precisely where a feature point is located. The standard deviation of the measurements is typically less than one pixel for the eye corners and filtering over time reduces the error to less than 0.5 pixels.

A brute force computation of such correlations is time consuming, since we work with fairly large kernels. For features such as eye or mouth corners the kernel size is typically 20x20 pixels. For identifying the chin or the whole mouth the kernel may be up to 50x25 pixels. In order to speed up the computation the correlation is implemented with line searches. The correlation function around the minimum tends to be smooth and can be approximated well by a parabolic function. For such a surface, the minimum is found quickly using conjugate gradient techniques.

The features we measure are mouth, chin, nostrils, eyes and eyebrows. Knowing the positions and shapes of these features is sufficient to identify visemes of the mouth and the most prominent emotional expressions. Sometimes the interior of the mouth is also analyzed to get a better measure of lip protrusion and stress put on the lips.

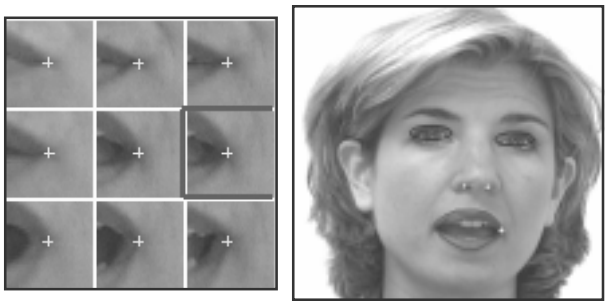

Figure 4: *Locating the corner of the mouth with a matched filter. From the nine samples on the left the corner belonging to a mouth with similar dimensions as the mouth on the right is chosen by the system. This kernel is scanned over the area around the mouth corner to find its precise location.*

3.4. Pose estimation

We compute the 3D pose of the head with the estimation technique reported in [8] using six feature points in the face: The four eye corners and the two nostrils. This technique starts with the assumption that all model points lie in a plane parallel to the image plane (corresponds to an orthographic projection of the model into the image plane plus a scaling). Then, by iteration, the algorithm adjusts the model points until their projections into the image plane coincide with the observed image points. The pose of the 3D model is obtained by solving iteratively the following linear system of equations:

$$\begin{cases} \mathrm{M_i} \bullet \dfrac{f}{Z_0} \mathrm{i} = x_i(1+\varepsilon_i) - x_0 \\ \mathrm{M_i} \bullet \dfrac{f}{Z_0} \mathrm{j} = y_i(1+\varepsilon_i) - y_0 \end{cases}$$

$\mathrm{M_i}$ is the position of object point i, i and j are the two first base vectors of the camera coordinate system in object coordinates, f is the focal length and $\boldsymbol{Z}_0$ is the distance of the object origin from the camera. i, j, and $\boldsymbol{Z}_0$ are the unknown quantities to be determined. x_i, y_i is the scaled orthographic projection of the model point i, x_0, y_0 is the origin of the model in the same plane, ε_i is a correction term due to the depth of the model point. ε_i is the parameter that is adjusted in each iteration until the algorithm converges. This algorithm is very stable, also with measurement errors, and it converges in just a few iterations.

4. Generating normalized face parts

Before the image samples are entered into the database they are corrected in shape and scale to compensate for the different head orientations when they were recorded. From the recognition module the position and shape of facial parts as well as the pose of the whole head are known. To extract facial parts from the images we first project the planes of the 3D model into the image plane. The projected planes then mark the extent of each facial part (Figure 5). These areas are warped into normalized bitmaps. Any information about the shape produced by the recognition module is also mapped into the normalized view and stored along the bitmap in a data-structure.

Once samples of a face part are extracted from the video sequences and normalized, they need to be labeled and sorted in a way that they can be retrieved efficiently. The first parameter used to describe a mouth shape is the phoneme sequence spoken during the recording. All recorded sequences are segmented into phonemes by a speech recognition system and the segmentation is double-checked by a human listener. A second set of parameters are the measurements produced by the recognition system. In figure 5, for example, we parameterize the mouth with three parameters: The width (the distance between the two corner points), the y-position of the upper lip (the y-maximum of the outer lip contour) and the y-position of the lower lip (the y-minimum of the outer lip contour). Samples of other facial parts are parameterized in a similar way.

Beside geometric features, we also use parameters describing the appearance of a facial part. The filtering processes described above provide a convenient way of characterizing the texture of a sample. By filtering a bitmap with a band-pass filter and measuring the intensity in seven frequency bands, we obtain a characterization of the texture that can be used to parameterize the samples. In this way, we can differentiate between samples that have the same geometrical dimensions, but a different visual appearance.

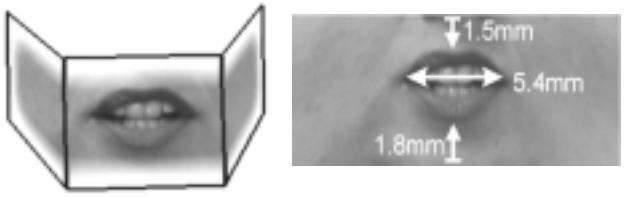

Figure 5: *The 3D model defines areas that are cut out for a facial part (left). The bitmaps are then normalized and the facial part is parameterized (right).*

5. Creating animated talking heads

A talking-head animation is driven by the output of a text-to-speech synthesizer (TTS). Starting from ASCII text plus some annotation controlling the intonation, the TTS produces a sound file. In addition it outputs a phonetic

transcription, including precise timing information for each phoneme plus some information about the stress. Given the sequence of phonemes and their durations, we search through the database for appropriate mouth shapes. A Vitterbi search calculates at each time step how well a mouth sample fits into the sequence. This evaluation takes into account the phonetic context, including how well the durations of the phonemes match those of the new sequence. Moreover, it compares the geometric parameters with those of mouth shapes coming before and after, to guarantee smooth lip movements. It also checks the parameters of visual appearance to make sure that there are no discontinuities in the appearance (see Figure 6).

We handle the animation of other facial parts using a model similar to the one developed for the MPEG4 facial animation subsystem [9]. Special markers are put in the text to control amplitude, length, onset and offset of facial animations.

5.1. Rendering

A frame of the final animation can be generated when bitmaps of all the face parts have been retrieved from the database. The bitmap of the base face is first copied into the frame buffer, and then the bitmaps of face parts are projected onto the base face using the 3D model and the pose of the base face. At the moment we consider only a limited range of rotation angles of ±10°, so there is no need for hidden surface removal. To avoid any artifacts from overlaying bitmaps, we use gradual blending or "feathering" masks. These masks are created by ramping up a blending value from the edges towards the center. These operations are implemented using basic OpenGL calls and the whole frame is rendered on the fly with just a few texture-map operations, which makes it possible to render the talking head in real time on a low cost PC.

6. Results and discussion

In order to obtain feedback on the quality of animated sequences, we conducted subjective tests with close to 200 people. One test evaluated whether a talking head could improve intelligibility of spoken text in a noisy environment. The head models improved intelligibility significantly, sometimes cutting the error rate by more than 50%.

Another test focused on the perception of a talking head in a user interface. People exposed to the head clearly preferred the interface with the head to a version without a head. Two thirds of the respondents judged the sample-based head as 'natural' or 'very natural'. The lip synchronization was judged mostly as 'good' or 'very good'. A complaint heard sometimes was that the heads are rather 'stiff'. Clearly, a lively animation is important to keep a viewer's attention.

7. References

[1] Parke, F.I., Waters, K., "Computer Facial Animation", A.K. Peters, Wellesley, Massachusetts, 1997.

[2] Guenter, B., Grimm, C., Wood, D., Malvar, H., Pighin, F., "Making Faces", Proc. of SIGGRAPH'98, pp.55-66, ACM SIGGRAPH, July 1998.

[3] Pighin, F., Hecker, J., Lichinski, D., Szeliski, R., Salesin, D.H., "Synthesizing Realistic Facial Expressions from Photographs", Proc. of SIGGRAPH'98, ACM SIGGRAPH, July 1998, pp.75-84.

[4] Bregler, C., Covell, M., Slaney, M., "Video Rewrite: Driving Visual Speech with Audio", Proc. SIGGRAPH'97, pp.353-360, ACM SIGGRAPH, July 1997.

[5] Ezzat, T., Poggio, T., "MikeTalk: A Talking Facial Display Based On Morphing Visemes", Proc. of Computer Animation, IEEE Computer Society, pp.96-102, June 1998.

[6] Cosatto, E., Graf, H.P., "Sample-Based Synthesi of Photo-Realistic Talking-Heads", Proc. of Computer Animation, IEEE Computer Society, pp.103-110, June 1998.

[7] Barron, J.L., Fleet, D.J., Beauchemin, S.S., "Performance of Optical Flow Techniques", Int. J. Computer Vision, Vol.12, pp43-77, 1994.

[8] Oberkampf, D., Dementhon, D., Davis, L., "Iterative Pose Estimation Using Coplanar Feature Points"; Internal Report, CVL, CAR-TR-677, University of Maryland, July 1993.

[9] Ostermann, J., "Animation of Synthetic Faces in MPEG-4", Proc. of Computer Animation, IEEE Computer Society, June 1998, pp.49-55.

Acknowledgement

We thank J. Ostermann, J. Schroeder and M. Potamianos for countless helpful discussions and comments.

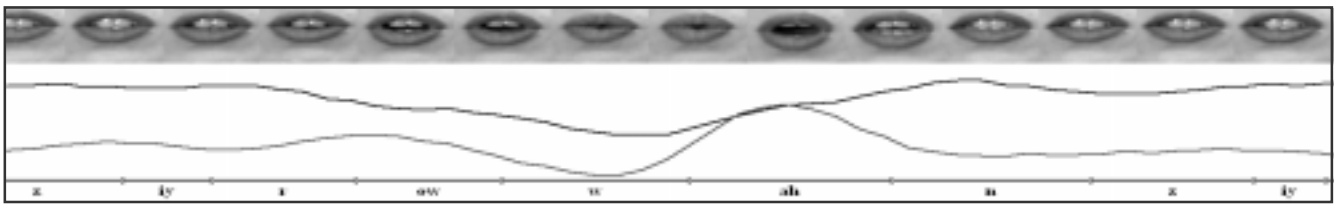

Figure 6: A synthesized sequence of mouth shapes. The bottom curve indicates the mouth height and the top curve the width. For the final animation these mouth bitmaps are warped to adapt to the head orientation of the base face.