Sample-Based Synthesis of Photo-Realistic Talking Heads

Eric Cosatto & Hans Peter Graf
AT&T Labs-Research, 100 Schulz Drive, Room 3-{124,134}, Red Bank, NJ 07701-7033, USA
{eric,hpg}@research.att.com

Abstract

This paper describes a system that generates photo-realistic video animations of talking heads. First the system derives head models from existing video footage using image recognition techniques. It locates, extracts and labels facial parts such as mouth, eyes, and eyebrows into a compact library. Then, using these face models and a text-to-speech synthesizer, it synthesizes new video sequences of the head where the lips are in synchrony with the accompanying soundtrack. Emotional cues and conversational signals are produced by combining head movements, raising eyebrows, wide open eyes, etc. with the mouth animation.

For these animations to be believable, care has to be taken aligning the facial parts so that they blend smoothly into each other and produce seamless animations. Our system uses precise multi-channel facial recognition techniques to track facial parts, and it derives the exact 3D position of the head, enabling the automatic extraction of normalized face parts.

Such talking-head animations are useful because they generally increase intelligibility of the human-machine interface in applications where content needs to be narrated to the user, such as educative software.

Keywords

Talking-head synthesis, sample-based synthesis, photo-realistic rendering, face recognition and location, sample-based coarticulation.

1 Introduction

Talking heads become more and more common as parts of modern computer-user interfaces. A talking head can add entertainment value to a program and make it more engaging. For example, children tend to listen more closely to an educative program, if a human face accompanies the narrating voice. Usually such heads are either recorded video sequences of real humans or cartoon characters that are synthesized in real time. Recorded video sequences are expensive to produce, require a lot of storage space, and limit the flexibility of the interface. Cartoon characters are suited for some applications because they can be entertaining, but often a photo-realistic talking head is more appropriate. Synthesizing a photo-realistic talking head is difficult, since our eyes are very sensitive to the slightest imprecision in modeling or rendering. Such artifacts are immediately noticed and will 'turn off' most users.

There are many other applications for photo-realistic talking heads, in particular in the area of very-low bit-rate coding of videos. The standards committee for the upcoming MPEG4 video compression standard introduced special provisions for encoding animated heads [1]. The goal is to transmit only a few parameter values, while the whole head is reconstructed at the receiver end from these parameters. To encode faces of humans in this way, an efficient synthesis of photo-realistic heads is needed.

Many 3D models of the human head have been developed [2]. Some even model the physical structures, such as bones and muscles, in great detail to derive the shape of the face [3][4]. Other 3D modeling techniques start with generic mesh models over which pictures of people are texture-mapped [7][8]. Recent efforts aim at personalizing 3D head models [5][6]. 3D models are very flexible for generating movements and showing the head in any desired orientation. However, to render highly deformable facial parts such as a mouth with a high degree of precision, complex models are needed that are compute-intensive and generally produce faces with a synthetic look.

An alternative approach is based on warping of sample views or morphing between several views [11][12][14][16]. These techniques can produce photo-realistic animations. The difficulty is, how to find the displacements of the image pixels. Warping a face requires precise specifications of the displacements of many points in order to guarantee that the results look like real faces. Most techniques therefore rely on a manual specification of the warp parameters. Poggio et al. [13][15] have proposed an image synthesis/analysis framework where the warp parameters are determined automatically, based on optical flow. While this approach gives an elegant

solution to generating new views from a set of reference images, one still has to find the proper reference images. Moreover, the computation of the "displacement maps", as well as that of the actual synthesis, are computationally expensive.

A synthesis technique based on recorded samples that are selected automatically has been proposed by Bregler et al. [9]. This system can produce video of existing persons (John F. Kennedy in one instance) uttering text they never actually said. Videos of triphones (3 subsequent phonemes) are used as the basic unit, resulting in a large library with tens of thousands of images. Despite the large number of samples, the system can not handle emotional expressions.

The technique described here is inspired by recent developments in automatic speech synthesis. Model-based synthesizers, developed over several decades, so far do not produce naturally sounding voices. More recently, most speech synthesizers striving for naturally sounding voice apply sample-based techniques, where short speech snippets are concatenated into new utterances.

In this paper we describe a similar method applied to video. We automatically extract samples from videos of a talking person and store them in a library. The number of samples needed for the synthesis is reduced by several orders of magnitude by decomposing the head into a hierarchy of facial parts. In this way individual parts, such as the mouth can be animated independently. This allows generating different emotional expressions while articulating text. The result is a compact system that can synthesize photo-realistic talking heads on a regular PC.

Section 2 describes the model of the talking head. Section 3 shows how to generate a compact library of sample bitmaps used for the synthesis. Section 4 explains how new sequences of a talking heads are synthesized, and how coarticulation is handled. In section 5 the process of image recognition is explained.

2 Model

In an image analysis/synthesis framework, we have to define similarities or distances between views of facial parts, in order to classify and retrieve them efficiently. For instance, in [9] the lip images are labeled with the sound that was produced while the lip was articulating speech. Similar sound patterns are assumed to generate similar lip shapes. This results in a sample library with many redundant images, because there are many more sounds than lip shapes. In our system we use measurements, such as lip width, lip opening, and jaw rotation, to describe the shape. This is more efficient, since now we can eliminate redundant views from our sample libraries. We use image recognition algorithms described in section 5 to locate facial features in images and measure their shapes precisely. This allows the automated capture of a large number of samples from real video footage and leads to a rich, yet compact library of samples from which photo-realistic animations can be synthesized.

2.1 Head and face parts

Separating a face into parts greatly reduces the number of different sample views needed for generating animated sequences. Consider adding a frown to a talking face. If the face is treated as a whole, a full set of mouth shapes, representing the articulation of all phonemes, has to be generated, together with the frowning eyes. If, on the other hand, eyes and mouth are treated as separate parts, we need only one set of mouth shapes for talking and two eye shapes, one for the neutral expression and one for the frown. An extra mouth shape may be added to emphasize the frown, when the mouth is closed. Adding more pronounced grooves may also underline the intended expression. Such variations are easy to achieve with a modular system, but would result in a combinatorial explosion in the number of stored expressions, if the whole face were treated as one unit.

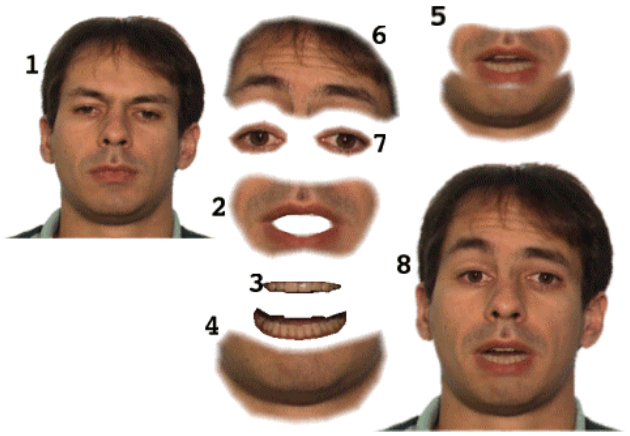

Figure 1*: Separating the head into facial part is essential for two reasons. First, it reduces the total number of parts needed to animate a talking head. Second, it reduces the number of parameters needed to describe a part. To generate a novel appearance, base head (1) is combined with mouth (5), eyes (7) and brows (6). The mouth area is generated by overlaying lips (2) over upper teeth (3) and lower teeth + jaw (4). This allows animating a jaw rotation independent of the lip shape.*

There is no unique way of decomposing a face into parts, and no part of the face is truly independent from the rest. Muscles and skin are highly elastic and tend to spread a deformation in one place across a large part of the whole face. The decomposition described here was chosen after studying, how facial expressions are generated by humans

[2][17], and how they are depicted by artists and cartoonists. Figure 1 shows a face and outlines the parts that form the model.

2.2 Visemes

A mouth shape articulating a phoneme is often referred to as 'viseme'. While over 50 spoken phonemes are distinguished in the English language, most researchers consider between 10 and 20 different visemes to be sufficient. We use twelve visemes, namely: a, e, ee, o, u, f, k, l, m, t, w, closed. All the other phonemes are mapped onto this set. In order to compare mouth shapes, we need to describe them with a set of parameters. For a first classification of the mouth shapes we choose 3 parameters: mouth width, position of the upper lip, and the position of the lower lip. The following table shows this parameterization for a few visemes.

viseme	sound	width	Position lower lip	Position upper lip
close		**0.5**	**0**	**0**
E	**f-ee-t**	**0.7**	**0.5**	**0.5**
U	**m-oo-n**	**0.2**	**0.5**	**0.5**
f	**f-eet**	**1.0**	**0**	**0.3**
…				

This representation is very convenient, since it defines a distance metric between mouth shapes and represents all possible shapes in a compact, low-dimensional space. Every mouth sample is mapped to a point in this parameter space. Now we can cover all the possible appearances of the mouth simply by populating the parameter space with samples at regular intervals.

This parameterization is somewhat crude and can not cover the finer nuances in the mouth shapes, i.e. two shapes with the same parameters may still look quite different to an observer. These details are handled by storing multiple copies of mouth shapes at each grid point in the parameter space (see section 3.2).

2.3 Head movement and emotional expressions

Head movements are an integral part of a talking head. Static heads, where only the lips move are usually not liked by users. For the synthesis of a talking head, the motions of all facial parts as well as movements of the whole head have to be planned carefully. The emotional state of the speaker is also signaled by changes in appearance of facial parts. For example, eyebrows lowered and drawn together indicate tension and fear [2][17].

A parameterization similar to the one for the mouth is used for the other facial parts, namely the eyes, the jaw, and the forehead. The eyes are allowed to blink, squint, and open widely. The iris can move left, right, up and down, the eyebrows can frown and raise, and the jaw moves up and down. Figure 2 shows a few examples of facial expressions that can be superimposed over regular speech to make it livelier.

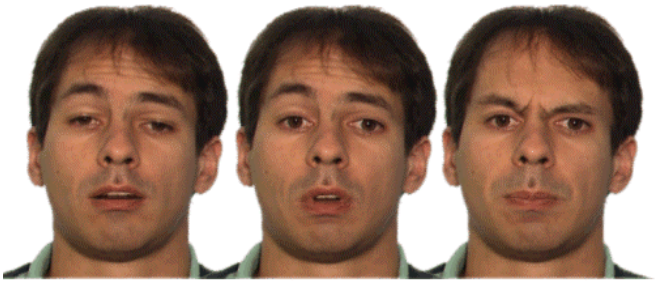

Figure 2: *These emotional expressions have been synthesized by blending independent face parts together. For example, sadness is expressed by raised eyebrows and semi-closed eyes (left), surprise by raised eyebrows and opened eyes (center), anger by lowered eyebrows and opened eyes (right). Such expressions are superimposed easily over speech since the lips are parameterized independently of the eyebrows and eyes.*

3 Sample extraction

Samples of face parts, such as the lips, are extracted from video sequences of a talking person. While talking, a head of a person is not stationary. Therefore, samples extracted from different frames of a video sequence must be realigned and reoriented so that measures made on one sample can be compared with the ones from other samples. Once realigned, parameters defining a given face part can be measured. Based on these parameters the samples are ordered in the libraries.

3.1 Viseme alignment

Using the image recognition techniques described in section 5, four points - two eyes and two nostrils - are located in the image of a face. The 2D coordinates of these four points are matched against successive projections of their corresponding 3D coordinates. These coordinates come from measurements on the subject itself, or can be approximated using generic anthropometric models [18]. Starting with the scaled orthographic projection and moving the object points along their line of sight, successive approximations of the projection are computed until convergence is achieved [19]. Then the projection of the mouth plane onto the image is computed using the head pose. Its position in the image defines a quadrilateral to quadrilateral mapping. The corresponding

transformation matrix is derived by solving a system of 8 linear equations [10]. The pixels are finally warped, as shown on figure 3, steps 1 and 2.

3.2 Populating the map

Once all the visemes are aligned with each other, their parameters are measured. The range of each parameter is determined and a quantization is defined, depending on the desired resolution. Intervals in the grid are not necessarily constant, but are smaller in areas of the parameter space where higher precision is desired. Then for each grid point, a search through all sample images is performed to find the closest viseme, that is, the one with measured parameters closest to the given grid coordinates. Figure 3, step 3 shows such a map populated with mouth visemes.

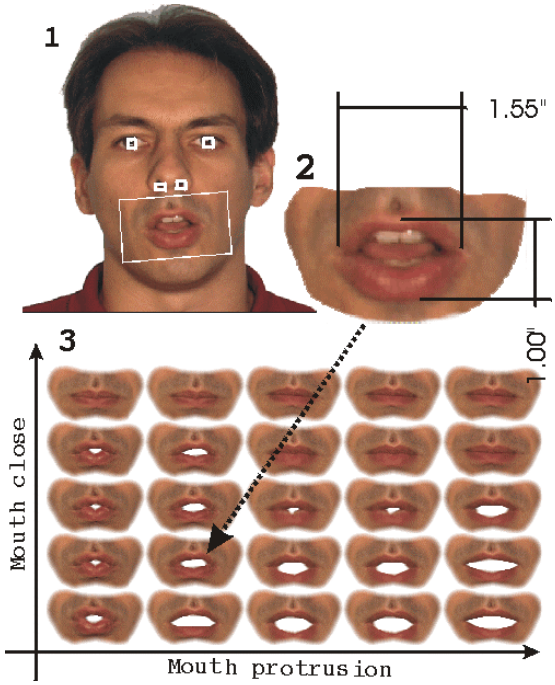

Figure 3: *From the 2D position of the eyes-nostrils plane in the image and the relative 3D position of these features in the "real world", the pose of the head is derived. The 2D position of the mouth plane is then derived from the head pose (1). Then the 2D mouth plane is warped and features can be measured (2). Once labeled each "viseme" is stored in a "grid" for indexing (3).*

A regular grid of visemes forms a hypercube in the space of parameters. Fully populating such a cube at regular intervals is not always possible. Some areas may not be covered, simply because the visemes don't exist (think of lips that are protruded, while fully open). After the parameter space has been populated automatically, the distribution of visemes is inspected visually. For that purpose we designed an interactive tool that allows to traverse the parameter space in arbitrary directions, showing these traversals as short animation snippets.

Often there are multiple visemes with the same lip parameters that look quite different to an observer. This is mainly due to different amounts of stress applied by the speaker. This is handled by allowing multiple bitmaps to occupy a grid point in parameter space and labeling each of them with a level of stress that is determined by an observer.

3.3 Effects of misalignment

Aligning the visemes correctly is critical, since slight misalignments lead to animations that appear jerky. These variations are not noticeable when looking at individual samples. However, they are quite disturbing when the samples are played in sequence. Since the visemes are collected from different video sequences, such variations can not be avoided completely. Our image analysis tools determine a facial feature's position within one pixel or less in most cases. However, image recognition sometimes makes mistakes and a viseme may end up misaligned. With our interactive inspection tool, we catch such mistakes easily and correct them.

4 Synthesis

A text-to-speech synthesizer (AT&T's Flextalk) provides the audio track of the animation as well as phoneme information at each time step. Figure 4 shows the block diagram of the whole system, including speech synthesizer and talking head generator.

4.1 Mouth synthesis and coarticulation

Naturally appearing mouth motion can only be achieved when coarticulation effects are taken into account. This means that the appearance of a mouth shape does not depend only on the phoneme pronounced at the moment, but also on the ones coming before and after. For instance, when we say "boo", the mouth shape for 'b' reflects that we plan to say 'oo'.

Our coarticulation model is similar to the one described by Cohen and Massaro [20]. We first define a mapping between each phoneme and the parameters defining the corresponding mouth shape. We also associate an exponential decay with each entry. Then, at each time interval, the mouth parameters are computed by accumulating the weighted influence of current, previous and following phonemes. In this way, a string of phonemes defines a trajectory in the space of mouth parameters. This

trajectory is then sampled at video rate (30 times per second). From each of these points, a mouth sample is generated by merging the two mouth samples with parameters closest to the sample point.

4.2 Sample-based coarticulation

We present briefly a novel idea for modeling coarticulation. As described in section 4.1, current coarticulation models are generic and do not account for individual ways of articulating speech. Our recognition system is able to track parameters of the mouth in video sequences of a person speaking and therefore enables extraction of characteristic lip motions. For this purpose we measure the lip parameters from a person speaking the most common tri-phones and quadri-phones and store them in a library. When a new string of phonemes is to be synthesized, a sliding window of three phonemes is scanned across this string. At each position, the tri-phone in the window is used to look up the mouth parameters of the center phoneme from the coarticulation library. In this way the coarticulation library translates sequences of phonemes into sequences of mouth parameters.

Since we need to store only the mouth parameters in the coarticulation library, even a large number of sample utterances fit in a compact library. The mouth parameters require a few bytes of storage, and with some compression the coarticulation library can be less than 100 kB in size.

At the moment we work with a relatively small coarticulation library (200 short sentences, about 1,000 triphones) that is generated from a database recorded for visual speech recognition [24]. Recording of a larger coarticulation library is under way.

4.3 Synthesis of other facial parts

Conversational signals are subtle movements of facial parts or of the head that emphasize a part of the speech, or regulate speech. Often a nod of the head accentuates a word, or a rising eyebrow indicates a question. Eye blinks also occur frequently and are usually synchronized with speech flow. Slight head movements generally accompany speech. When such motions stop, it often means that the speaker has finished and is expecting the listener to take action.

We add movements of the head and the facial parts that are in part random, and in part depend on the state of the talking head. We use a number of movement patterns that are representative for movements observed in speakers when we recorded the sample databases. Markers can also be put directly into the input text to trigger desired movements of facial parts at a given time during the animation. In this way, if a user knows what emotions should be expressed, this can be added explicitly by annotating the text.

Samples for the secondary facial parts are recorded and labeled using the same techniques as described for the mouth. For example, movements of the eyebrows are recorded and labeled by measuring their positions relative to the eyes. Then, for the synthesis the desired samples are recalled from the library based on the state of the talking-head.

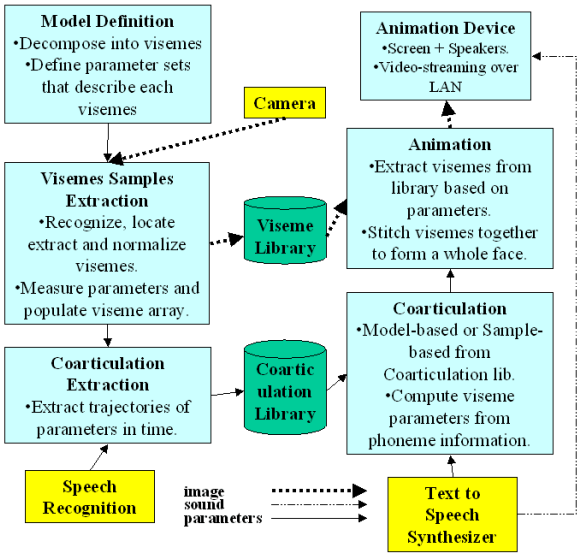

Figure 4: *System overview. The left side shows how a model is generated, while the right side shows the animation path.*

4.4 Final animation

To provide head movements, whole heads are recorded and labeled using the 3D head pose detection technique described in section 3.1. Movements are generated according to the state of the talking head plus some random motions. Once a base head has been selected, its pose is used to warp the facial parts onto it. All face parts are alpha blended using a feathering mask to avoid boundary artifacts.

5 Recognition of facial parts

Key to a sample-based synthesis approach is a reliable recognition of facial parts in images. This is needed in order to populate the library of sample parts. A small number of samples can be extracted manually. Yet such an approach becomes impractical for lively animations where thousands or even tens of thousands of images have to be searched in order to find appropriate samples. For sample images to be integrated into an animation, their positions

have to be determined very accurately, namely, within one pixel or even better.

Finding the location of a facial feature requires the discrimination of fine nuances in texture and color. For example, often there is hardly any contrast between the lips and the surrounding skin, which makes it difficult to locate the outlines of the mouth precisely. Even less clearly defined are such features as the lower edge of the chin. Moreover, when a speaker pronounces plosives, for example 'b' or 'p', the lip motion is so fast that mouth shapes in subsequent frames, recorded at 30 Hz, can differ substantially. One can therefore not rely on simple tracking algorithms that assume finding a shape similar to the one in the previous frame.

Such problems led many researchers to have the speakers wear marker points or lipstick, which provides a high contrast with the surrounding skin. Some people use head-mounted cameras to maintain constant geometric arrangements. While such intrusive techniques tend to provide measurements with a high accuracy, they disturb speakers, often to a point where their movements are no more natural.

Recently a surge of interest in recognizing faces has led to a number of algorithms for locating facial features. In particular, measuring mouth shapes has received a lot of attention from the speech reading community [21][22]. While there has been great progress, such algorithms still tend to be rather brittle. They work under well-controlled conditions, but often fail when the speakers or the lighting conditions change.

Our system uses multiple channels of analysis, namely: shape and texture analysis, color analysis and motion analysis. Having multiple channels of analysis increases the robustness considerably. It has been tested on a database of fifty different speakers with a total of over 200,000 frames. The mouth shapes could be extracted successfully for all these subjects, regardless of the complexion of the speakers or whether beards or moustaches were present. We give here a brief overview of the recognition system; more details can be found in [23].

5.1 Image analysis

In a first step the whole image is searched for the presence of heads, and their locations are determined. Then we zoom in on particular facial features to analyze them in more detail. Regardless of whether whole heads or individual facial features are being investigated, the image analysis proceeds in the same way, as shown schematically in Figure 5.

The video stream is processed in three separate channels. The first channel takes monochromatic images as input and searches for the presence of certain shapes and textures. In the second channel color segmentation is done, and in the third one the motion is estimated based on frame differences.

5.2 Representation of the data

Each of the channels produces a set of features and combinations of these features are evaluated with classifiers. Since features produced in different channels may have different representations, it is necessary to provide a scheme to compare different representations. They include bitmaps cut from the image after band-pass filtering, binary bitmaps, splines marking the outlines of an area, and simple geometric shapes such as a bounding box or a single point marking the position. In order to save compute-time, the analysis starts with simple representations, and only if the result is not satisfactory, a more complex representation is used. For example, when a classifier tries to determine whether three features represent two eyes and a mouth, it takes in a first pass only the center of mass of each feature into account and measures their relative positions. In a next step the

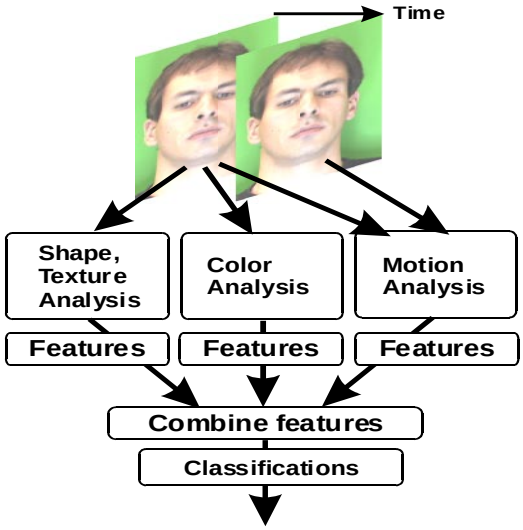

Figure 5: *Schematic overview of the image analysis process.*

classifier also looks at the shape of each feature, at which point the outline representation or, if available, the binary bit map is used.

For each of the representations a distance metric is defined to measure similarity between shapes. The distance metrics are defined between identical representations as well as between different ones.

5.3 Combining information

Each of the channels of analysis produces shapes in one or several of the representations described above.

Individual shapes carry little information and may mark elements that have nothing to do with a face. Only the combinations of several such shapes are useful to decide whether a face is present. Combining shapes is done with an 'n-gram' search. First each shape is analyzed individually. Those totally out of range in size or aspect ratio can definitely not represent a facial feature and are discarded. Then each of the remaining shapes is labeled with the facial features it might represent, e.g. eye, nostril, head edge, etc.

Next, combinations of two shapes are tested, whether they can represent a pair of facial features, for example an eye pair, eyebrows, or an eye plus a mouth. In the next step triplets are evaluated, etc. In each of these steps the geometric arrangements of the features are evaluated with

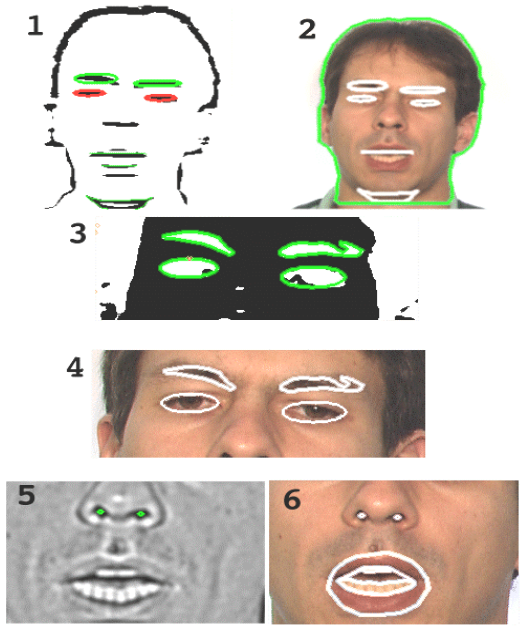

Figure 6: *Examples of processing steps to identify facial features. 1 shows an intermediate step of the shape analysis used to locate the head and give an estimate of the positions of a few facial features as shown in 2. 3 shows a result of the color segmentation to locate eyes and eyebrows as shown in 4. 5 illustrates an intermediate step of the shape analysis, used to locate nostrils and mouth. 6 shows the results of shape and color analysis, locating outer and inner edges of the lips, the teeth and nostrils.*

small classifiers that take as their inputs the sizes, ratios of distances, and orientations of the shapes. If information about the reliability of the different channels of analysis is available, the features are multiplied with a weight factor before entering into the classifiers.

The system for locating facial features consists of libraries of algorithms that are integrated into a particular application with a script language. When generating the samples for the viseme library, one typically deals with a single person and lighting conditions that give an even illumination of the face. Locating the head is a trivial task under such circumstances. However, the locations of the facial features, such as the mouth, the eyes, and the eyebrows, have to be determined very precisely. This is a challenging problem even under these conditions. The algorithms typically proceed with the shape analysis first to obtain a rough idea of a facial feature's position. Then the colors in this area are calibrated with a leader-clustering algorithm, followed by color segmentation. Information about motion is used only as an indicator whether a fast lip movement is taking place or whether the head is moving.

Figure 6 shows a few results of the feature location process. For the viseme library discussed here the speaker was recorded pronouncing 50 different utterances that cover the most prominent transitions of phonemes. This suffices to fill most of the parameter space of visemes. There are a few 'holes' left that have to be filled with morphed samples.

In addition to the single-speaker recordings for the viseme library we have analyzed a library of 50 different speakers with a total of 2,500 utterances, where each utterance consists of a video with 80 to 110 frames [24]. These samples were selected for lip reading experiments, yet are also well suited to compare coarticulation in different speakers. A subset of this database is used at present as the coarticulation library.

When the recognition system is trained with a particular speaker, the recognition accuracy for finding the mouth location is higher than 98%. The location of a facial feature usually can be determined within one pixel or better. To achieve such an accuracy, smoothing with linear interpolation over five to ten frames is done. For measuring the center of the mouth or the eyes that is no problem, if a face is recorded at 30 frames per second. Only the motion of the lips is often too fast that such techniques can not be applied for locating lip edges.

6 Conclusions

In this paper, we have presented a system that can animate a *photo-realistic* talking head. The method is based on image samples, allowing fast rendering by simple image overlay. The key differences with flip-book techniques are two-fold. First, the model of the head is composed of face parts that can be animated individually, thereby reducing the total number of required images by several orders of magnitude. Second, each of these facial parts is labeled with measured features that have a semantic meaning, such as "lip width". We use robust

image recognition techniques to locate and measure automatically face parts in video sequences of a talking person. Labeling of the face parts allows extracting a compact, representative set of appearances of each face part. We use a set of lips that cover speech postures, and we use other face parts such as eyes and brows to convey emotions. A speech synthesizer producing a stream of phonemes drives the animation. Using coarticulation the stream of phonemes is mapped into a smooth trajectory in the space of visemes. By sampling this trajectory at video rate, an animation is created for each face part. A frame of the final animation is composed of a base-head onto which each face part is warped and blended using a transparency mask.

6.1 Discussion and Future Aspects

There are several possible ways of generating animated talking heads, and the preferred method depends on the specific requirements of the application. The strengths of 3D head models and those of sample-based techniques are complementary to a large extent. In an environment with a limited number of movements and views a sample-based approach looks promising. It can deliver photo-realism that is still hard to match with texture-mapped models. However, if the emphasis is on a wide range of motions, the greater flexibility makes a 3D model advantageous. Maybe in the future a combination of the two techniques can combine the best parts of the two solutions.

Future directions include the use of a generic 3D model onto which samples can be texture mapped. This would increase the flexibility of the synthesis and would allow richer movements. To maintain a photo-realistic appearance sample views from all sides of the head are then required.

References

[1] J.Ostermann, A.Puri, **Natural and Synthetic video in MPEG-4**, Proc. of ICASSP 1998.

[2] Frederic I. Parke, Keith Waters, **Computer Facial Animation,** A.K. Peters, Wellesley, Massachussets, 1997.

[3] N. Magnenat-Thalmann, P. Primeau, D. Thalmann, **Abstract Muscle Action Procedures for Face Animation**, Visual Computing. 3, 1988, 290-297.

[4] D. Terzopoulos, K. Waters, **Analysis and Synthesis of Facial Image Sequences Using Physical and Anatomical Models**, IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 15/6, pp. 569-579, 1993.

[5] J. Ostermann, L.S. Chen, and T.S. Huang, **Animated Talking Head with Personalized 3D Head Model**, to be published 1998.

[6] M. Escher and N.M. Thalmann, **Automatic 3D Cloning and Real-Time Animation of a Human Face**, Proceedings of IEEE Computer Animation 97, pp. 58 - 66.

[7] S. Morishima, H. Harashima, **A Media Conversion from Speech to Facial Image for Intelligent Man-Machine Interface**, IEEE J. Selected Areas in Communications, 9, 1991, pp. 594-600.

[8] T. Kuratate, F.Garcia, H.Yehia, E.Vatikiotis-Bateson, **Facial Animation from 3D Kinematics**, ASJ, Sept 1997, Sapporo.

[9] C. Bregler, M. Covell, M. Slaney, **Video Rewrite: Driving Visual Speech with Audio**, Proc. SIGGRAPH'97, pp.353-360.

[10] G. Wolberg, **Digital Image Warping**, IEEE Computer Society Press, Los Alamitos, CA, 1990.

[11] S.M. Seitz, and C.R. Dyer, **View Morphing**, Proc. SIGGRAPH 96, pp. 21-30.

[12] N. Arad, N. Dyn, D. Reisfeld, and Y. Yshurun, **Image Warping by Radial Basis Functions: Applications to Facial Expressions**, CVGIP: Graphical Models and Image Processing, 56, 1994, pp. 161-172.

[13] D.Beymer and T.Poggio, **Image Representation for Visual Learning**, Science, 28 June 1996, vol 272, pp1905-1909.

[14] T. Beier, and S. Neely, **Feature-Based Image Metamorphosis**, Computer Graphics, 26, 1992, pp. 35-42.

[15] T. Ezzat and T. Poggio, **Facial Analysis and Synthesis Using Image-Based Models**, Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition, IEEE CS Press, 1996, pp.116-121.

[16] M. Bichsel, **Automatic Interpolation and Recognition of Faces by Morphing**, Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition, IEEE CS Press, 1996, pp.128-135.

[17] P. Ekman, and W. Friesen, **Facial Action Coding System: A Technique for the Measurement of Facial Movement**; Consulting Psychologists Press, Palo Alto, CA, 1978.

[18] T.Horprasert, Y.Yacoob, L.Davis, **Computing 3D Head Orientation from a monocular image sequence**; Proc. 2nd Int. Conf. on Automatic Face and Gesture Recognition, IEEE CS Press, 1996, pp.242-247.

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

[20] M.M. Cohen, D.W. Massaro, **Modeling Coarticulation in Synthetic Visual Speech**, in: Models and Techniques in Computer Animation, M. Magnenat-Thalmann and D. Thalmann (eds.), Tokyo, 1993, Springer Verlag.

[21] D.G. Stork and M.E. Hennecke (eds.), **Speechreading by Humans and Machines**, Springer, Berlin, 1996.

[22] Proc. Audio-Visual Speech Processing, C. Benoit and R. Campbell (eds.), Rhodes, Greece, 1997.

[23] H.P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan, **Multi-Modal System for Locating Heads and Faces**, Proc. 2nd Int. Conf. Automatic Face and Gesture Recognition, IEEE Computer Soc. Press, 1996; pp. 88-93.

[24] G. Potamianos, E. Cosatto, H.P. Graf, and D.B. Roe, **Speaker Independent Audio-Visual Database for Bimodal ASR**, Proc. Audio-Visual Speech Processing: 1997, pp.65-68.