VISUAL SPEECH SYNTHESIS FROM 3D VIDEO

J.D. Edge and A. Hilton
Centre for Vision, Speech and Signal Processing,
School of Electronic and Physical Sciences, University of Surrey, Guildford, GU2 7XH, UK.
j.edge@surrey.ac.uk, a.hilton@surrey.ac.uk

Keywords: Facial Animation, Speech Synthesis, Virtual Humans.

Abstract: Data-driven approaches to 2D facial animation from video have achieved highly realistic results. In this paper we introduce a process for visual speech synthesis from 3D video capture to reproduce the dynamics of 3D face shape and appearance. Animation from real speech is performed by path optimisation over a graph representation of phonetically segmented captured 3D video. A novel similarity metric using a hierarchical wavelet decomposition is presented to identify transitions between 3D video frames without visual artifacts in facial shape, appearance or dynamics. Face synthesis is performed by playing back segments of the captured 3D video to accurately reproduce facial dynamics. The framework allows visual speech synthesis from captured 3D video with minimal user intervention. Results are presented for synthesis from a database of 12minutes (18000 frames) of 3D video which demonstrate highly realistic facial animation.

1 INTRODUCTION

In recent years the use of data-driven techniques to produce realistic human animations have become more prevalent. The use of video and motion-capture data allows us to improve the visual realism of a synthetic character without the complexities involved in physical simulation. Unfortunately, whilst video data is lifelike, it is also constrained to the 2D image plane. Furthermore motion-capture technologies are marker-based and so prevent the recovery of skin texture during data capture. 3D video capture technology uses stereo-reconstruction techniques to capture both the texture and geometry of the facial surface thereby providing a best compromise solution to animating realistic human characters.

One of the most difficult problems in animation is reproducing the visible aspects of speech. It is the method of communication that we use every day, and any viewer will instantly spot any disparity with everyday reality. When we see someone speaking the movements of the articulators (tongue, jaw, lips etc.) is actually creating the sounds that we hear, and we know there is a high correlation between what is seen and heard (e.g. lip-reading (Sumby and Pollack, 1954) and the McGurk effect (McGurk and MacDonald, 1976).) Whereas traditional animation techniques are adequate for creating cartoon-like speech, we need more complex models to accurately synthesize realistic human speech movements. Data-driven techniques are particularly appropriate to this problem because they remove the need to directly model the dynamics of speech by retrieving this information from a real speaker. In this paper we introduce a data-driven technique which uses 3D video capture technology to animate realistic speech movements.

Our process of producing a talking head works on a similar basis to Video-Textures (Schödl et al., 2000). We capture sequences of 3D speech movements and organise this data into a graph structure consisting of nodes representing dynamic phonetic units with connecting arcs representing optimal transitions. Speech synthesis is achieved by traversing this graph according to the phonetic structure of an input utterance. Animations consist simply of playing back frames of the original captured data with no blending/interpolation required. Our process of creating a talking head has been designed to as far as possible remove the necessity for any manual intervention or reliance upon sensitive computer vision algorithms.

2 BACKGROUND

Visual speech synthesis is complicated by the fact that when producing sequences of speech sounds, the action of the articulators in the vocal tract are greatly affected by speech context. This context-dependency of speech is often termed 'coarticulation' (Löfqvist, 1990). The physical phenomena occurs bidirectionally, that is articulation is affected by both preceeding and upcoming speech movements. Furthermore, certain aspects of speech are more important than others; for example position of the tongue whilst producing a /θ,ð/ (as in *th*in or *th*is) sound exhibits low variation and is therefore a highly dominant feature, whilst jaw rotation for vowel sounds exhibits high variation and has little affect on surrounding speech segments. The simulation of coarticulation is highly important for the naturalness of any speech synthesis technique (both audible and visual.) Typically, coarticulation is simulated using a form of spline blending of articulatory parameters (Ezzat et al., 2002; Cohen and Massaro, 1993).

More recently, captured dynamics have been directly used in the synthesis of visual speech. This mirrors the most common method of audible speech synthesis which takes captured audio data and concatenates small sections to form novel utterances (Taylor et al., 1998). In Video-Rewrite (Bregler et al., 1997) triphone sequences of video data are concatenated to synthesize lip movements which are then pasted onto a background sequence of head movement improving the naturalness of the output animation. In (Kalberer and Van Gool, 2002) a model of speech movements is captured using markers, then a viseme-space is constructed using statistical methods and navigated to synthesize new utterances. In (Cao et al., 2003; Kshirsagar and Magnenat-Thalmann, 2003) motion-captured facial movements are used to drive a 3D model-based synthesis technique. These techniques inherit from earlier work on motion graphs for body motion capture data (Kovar et al., 2002).

The main difference between the audio and visual concatenative methods is that it is more difficult to capture facial movement than audio. Whilst facial dynamics have traditionally been captured using markers, recently surface capture techniques have been developed which provide greater spatial resolution in 3D (Wang et al., 2004; Ypsilos et al., 2004; Zhang et al., 2004). Surface capture has the added benefit that surface texture can be simultaneously captured - providing fine detail of skin wrinkles and creases that will always be invisible to marker-based technologies. In our work surface capture technology is used to provide data for speech synthesis.

3 OUR APPROACH

The technique for visual speech synthesis introduced here relies upon the concatenation of small phonetic units representing the variation in the dynamics of natural speech movements. Our approach can be summarised into several phases:

- Data Capture (Section 3.1) - A database of natural speech utterances are captured from an actor. These utterances contain the variations, due to coarticulation, in speech articulation. The speech data consists of 3D video and phonetically labelled audio of a news corpus.
- Graph Construction (Section 3.3) - The speech data is split into phonetic units, and transition probabilities between all units in the captured corpus are calculated. This leads to a graph structure, subsets of which are traversed during synthesis. Transition probabilities are related to the similarity of frames in different phonetic units.
- Unit Selection (Section 3.4) - For a given output utterance, split into its phonetic constituents, appropriate units (dynamic phonemes) are selected from the constructed motion graph. Selection is performed using a Viterbi algorithm which maximises the probability of a sequence of units which match the phonetic structure of the output utterance.
- Animation (Section 3.5) - Selected phonetic units are resampled and played back with the audio to animate the speech utterance. No interpolation or processing of the frames is performed, the final animation is simply a re-ordering of frames from the initial speech corpus.

Our system has been designed to minimise the manual intervention required to create new talking heads. All the above stages beyond data capture can be performed automatically, and the only intervention currently required is to correct the automatic phonetic labelling of captured audio.

3.1 Data Capture

All data-driven techniques require a significant initial data capture to represent the range of possibilities in synthesis. In our system the data capture consists of words and sentences from a news corpus spoken by an actor. The actor is captured using a custom 3D face capture rig (see (Ypsilos et al., 2004)), which reconstructs facial geometry (200×200 scalar elliptical depth maps) and texture (512×512 RGB images). Audio is captured simultaneously and phonetically labelled to facilitate the construction of a motion graph.

In total 12 minutes of audio/video data has been captured to drive our talking head. It is most important to capture a large number of consonant units, as these units visually structure speech whereas vowels are usually transitional movements. It is also the case that vowels may to a large degree be interchangeable, although this is not currently implemented in our system. Other than voiced/voiceless contrasts we maintain all phonetic distinctions in the database, mainly because a viseme reduction is less valid when the units themselves are dynamic (i.e. the central frame may be the same, but the articulatory movement may exhibit a high degree of variation even within a viseme class.)

3.2 Similarity Metric

In order to optimally construct a transition graph of the 3D video data a similarity metric is required to compare individual frames. Each frame consists of both geometry and texture, which should both be taken into account when determining similarity ($F_i = \{G_i, T_i\}$). This section describes the creation of a similarity metric between 3D video frames, i.e. $dist(F_i, F_j) \rightarrow \mathbb{R}$.

In (Schödl et al., 2000) a simple $L2$ similarity metric is used to recover transitions between 2D video frames, however we have found that the effect of noise (e.g. sensor noise, geometry reconstruction error) upon this metric is highly detrimental and leads to poor output transition graphs. This is due to the high dimensionality of each frame in the original data ($(200 \times 200) + (3 \times 512 \times 512) = 826432$ scalar values per frame). In (Ezzat et al., 2002) principal components analysis (PCA) is used as a dimensionality reduction technique on 2D video, unfortunately given the size of the database used here (18000 frames $\times 826432$ scalar values) makes traditional PCA techniques unviable. It is also important to note that traditional PCA applied by compressing each 2D image frame into a single vector does not optimally account for spatial redundancies in the data (Shashua and Levin, 2001). To make up for these limitations we apply a hierarchical variant of PCA which accounts well for both spatial and temporal redundancies in the data and is more efficient to create with very large databases. The projection into this low dimensional space is used to provide a similarity metric between 3D video frames.

A wavelet decomposition of a signal (e.g. a row/column of image data) is a frequency representation produced by projecting onto a wavelet basis (e.g. Haar, Daubechies etc.) The decomposition consists of a signal average V^0 and sets of wavelet coefficients W^m. Any level of the frequency hierarchy ($V^0 \subset V^1 \subset V^2 \subset \ldots V^m \ldots \subset V^n$) can be reconstructed using the W^m and scaled/shifted basis functions ψ^m. If a sequence of images is decomposed into wavelet coefficients, we can apply PCA to the W^m to recover a model of how different spatial frequencies change over time.

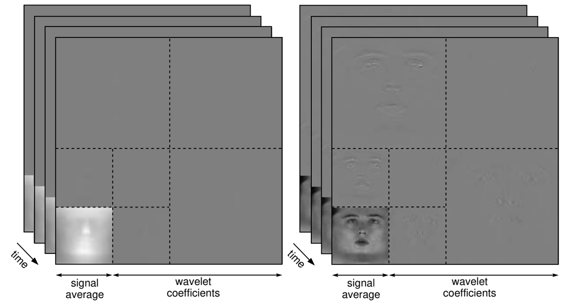

Figure 1: Each frame (both texture and geometry) in the database is converted to wavelet coefficients, and the principal components of these coefficients over time are used to define a similarity metric. This examples show only two levels of the wavelet decomposition.

$$W^m = \mu_{W^m} + \sum_k (b_k^m . V_k^m) \qquad (1)$$

In (1), μ_{W^m} is the mean of the W^m coefficients over time, the V_k^m are the principal components representing the important variations of the W^m over time. The b_k^m give a projection onto the recovered principal components, and act as the parameterisation of the original image sequences. The proposed wavelet/PCA decomposition explicitly localises variation spatially (i.e. the wavelet transformation) and temporally (i.e. the principal components). This wavelet/PCA combination is demonstrated in fig. 1.

In practice we decompose each frame (both texture and geometry) using a Haar wavelet basis in an alternating row/column scheme (the non-standard approach described in (Stollnitz et al., 1995).) A $3 \times 3 \times 3$ (rows $\times$ columns $\times$ frames) spatio-temporal median filter is applied to the highest-frequency components of the wavelet transform, which is used to attenuate noise spikes in the data. From the projection onto the V_k^m (the principal components across all frequencies) 10 components are selected which account for over 95% of the variance in the data, this is the basis used for comparison. The similarity metric, $dist$, is defined as the L2 distance between frames projected onto the selected basis. We do not use the Mahalanobis distance, as used in (Ezzat et al., 2002), because this gives too much weight to low variance components which are relatively insignificant when determining similarity.

The similarity of any two frames in the database can now be determined using the selected basis. This distance can now be converted into a frame-to-frame transition probability.

$$P(F_i, F_j) = \exp\left(\frac{-dist(F_i, F_j)}{\sigma}\right) \qquad (2)$$

In (2) σ is a constant which relatively scales the probability of transitioning between two frames (in our system set to half the average distance between frames, μ_{dist}.) This allows the construction of a matrix containing transition probabilities between all frames in the stored data. In a final step we apply a small (3 frame window) diagonal filter to the data similar to (Schödl et al., 2000), this incorporates a notion of derivative similarity in the transition probability. This transition matrix is used in the construction of a transition graph between dynamic phonetic units.

3.3 Graph Construction

The synthesis of speech movements is based upon the traversal of a graph representing the phonetic structure of the captured data. For these purposes phonemes are considered to be sequences of frames from the centre of the previous phoneme to the centre of the following phoneme. Furthermore, we merge voiced/voiceless contrasts in our phoneme units (e.g. /p/ and /b/ are considered to be the same unit), which maximises the available dynamic units available in the database. It is important to note that we only merge voiced/voiceless consonant contrasts; often nasals are also combined, e.g. /m/ is often merged with the /p,b/ group, yet these are not dynamically the same and so this is not advisable in a concatenative system.

The units are similar to what are often considered as triphones (e.g. in (Bregler et al., 1997)), however, only the central phonetic label of the unit is used during synthesis (i.e. no phonetic context is explicitly taken into account) thus we refer to these as dynamic phonemes. This is an important distinction because triphone synthesis typically requires many more units to be captured. By disregarding the previous and following phonetic labels, and matching only according to similarity, we can maximise the use of smaller speech databases.

Each phonetic unit consists of two periods, the onset (between the start of the unit and its phonetic centre) and the offset (between the phonetic centre and the end of the unit.) Possible graph transitions $\overrightarrow{q_i q_j}$ between phoneme states q_i and q_j occur between frames in the offset of q_i and frames in the onset of q_j. There will only be one optimal 'stitching' point between the two sequences, the point at which the frame-to-frame transition probability (see Section 3.2) is highest, therefore we have a directed graph with transition probabilities between each discrete state.

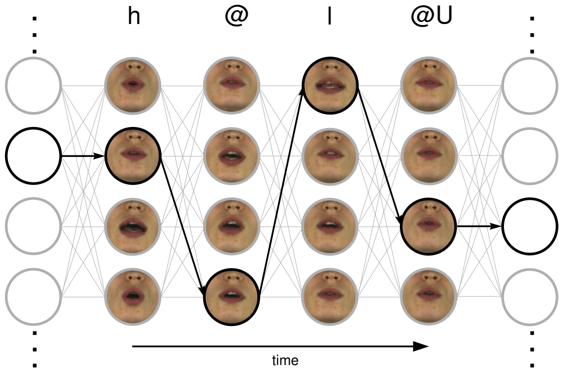

Figure 2: A graph of nodes representing dynamic-phonemes is constructed, each node is connected to each of the following phonetic units in the sequence, the Viterbi algorithm is used to find the least expensive path (shown in bold) matching an input utterance.

In the data set we have captured there are 1314 phonetic units, from a set of 38 British English phonemes, giving 1314^2 possible transitions. By thresholding out low probability transitions and pruning the graph according to phonotactic rules (e.g. bilabial plosives cannot occur sequentially) and removing within-state transitions the search space to be traversed during synthesis can be greatly reduced. The number of phonemes represented in the graph could potentially be reduced further, as the main difference between many speech sounds is the position of the tongue which is virtually indistinguishable in our data (except in some key examples such as /θ,ð/, e.g. *th*ing.) However, we did not follow this route because often the dynamic properties of phonemes are distinct even when the central lip pose is very similar. The size of our speech database allows us to represent phonemes and not resort to a coarse viseme reduction.

3.4 Unit Selection

The constructed phonetic graph structure allows synthesis to be formulated as a Markov process, i.e. the transition between phoneme states is independent of previously traversed states (3). This is obvious, given that the probability of transitioning between any two phoneme states in the graph depends only upon the similarity of frames (as defined in Section 3.2.) Thus, given a sequence of phoneme states (e.g. 'good evening' → /ɡ/,/ʊ/,/d/,/iː/,/v/,/n/,/ɪ/,/ŋ/) we must select the best sequence of states from the constructed graph (see fig. 2) which visually represent these speech sounds. One of the most common methods for doing this is the Viterbi algorithm.

$$P(q_i,\ldots,q_n) = \prod_{i=1}^{n} P(q_i|q_{i-1}) \qquad (3)$$

The Viterbi algorithm is a recursive solution to finding the best sequence of states matching a series of observations (i.e. the phonemes.) Given the Markov assumption (3), we can calculate the best probability from any of the possible initial states q_0 optimally by determining best partial paths to intermediate q_i states which are reused in the calculation of the q_{i+1} transition. This is a fast method of finding the best n states to match the phonetic input.

For the previous example ('good evening') each of the phonemes is a state in the Viterbi search, with m possibilities (where m is the number of examples of that phoneme in the database.) So each state may have a different number of alternatives, as the database does not have a uniform sampling of all phonemes. Therefore the probability of a sequence, $P(\text{ɡ}_5,\text{ʊ}_3,\text{d}_6)$ will be the product of the transition probabilities $\overrightarrow{\text{ɡ}_5\text{ʊ}_3} \times \overrightarrow{\text{ʊ}_3\text{d}_6}$ (where $\overrightarrow{xy}$ is the scalar transition probability from $x \rightarrow y$). The Viterbi algorithm simply reformulates this potentially exhaustive search into a series of subproblems, e.g. $\{\text{ɡ}_{best}\{\text{ʊ}_{best}\{\text{d}_{best}\{\text{iː}_{best}\{\text{v}_{best}\{\text{n}_{best}\{\text{ɪ}_{best},\text{ŋ}_{best}\}\}\}\}\}\}\}$, where the innermost problem is tackled first and recursively the inner problems accumulate to construct the best path from all possible states.

Note that audio parameters (e.g. cepstral coefficients) are not used to optimize the chosen states (e.g. as seen in (Cosatto and Graf, 2000)), this could be added as a further term to allow animation directly from speech audio with no intermediate phonetic transcription. We find that optimising on the transition probabilities alone provides smooth transitions which is the most important factor in creating high quality animation. A further advantage being that where sequences of phonemes in the target utterance are present in the training data they will be selected as a sequence (because the transition probability between sequential units will be 1). The unit selection procedure intentionally biases toward long contiguous sequences in the original captured data, which will maximize the quality by reducing the number of synthetic jumps. Note that the Viterbi graph search is performing a similar role to the greedy tile-matching procedure used in (Cao et al., 2003), but will not choose long sequences if they have a high associated transition cost with surrounding units. Tile-matching biases toward longest units at the expense of possible poor transitions because boundary matching has no associated cost (except where there are two competing units of the same length.) We would assert that the Viterbi algorithm is a better all around solution.

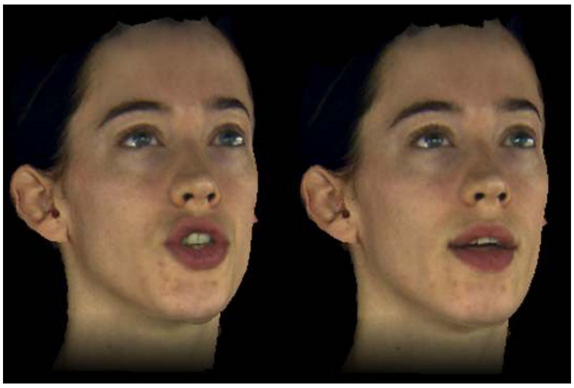

Figure 3: Frames from a synthetic utterance.

3.5 Animation

Our system only requires a rudimentary treatment of animation. Frames from a synthetic sequence can be seen in fig. 3. Since our raw data is 3D video of an actors face there is no necessity for complicated facial models or deformation techniques. The output of unit selection is a sequence of dynamic phonemes (i.e. sequences of frames from our initial data) and relative timings with the input utterance (i.e. the disparities between phoneme centres in the input utterance and our dynamic phonemes.) We use simple linear scaling to align selected units correctly with the utterance audio, and where the stretching/squashing of the dynamic phonemes leads to a misalignment of frames in the two sequences the closest frame is chosen (i.e. no interpolation of frames.) The only intervention in the final animation is to use a noise function to apply random low-magnitude rotations to the head with respect to the neck. This prevents the synthetic character from appearing too static, as can be seen in many 2D video-based talking heads.

Animations created using our system show several advantages over traditional mocap techniques (Cao et al., 2003; Kshirsagar and Magnenat-Thalmann, 2003). By capturing the surface texture of the actor simultaneously with the geometry we can recover and display high resolution features which would be lost using techniques such as motion-capture. It is also noticeable that that idiosyncrasies particular to our speaker are captured, an obvious example for our actor is the asymmetrical way that she opens her mouth. The quality of animations generated using our synthesis technique are on a similar to those produced from 2D video (e.g. (Ezzat et al., 2002; Brand, 1999; Bregler et al., 1997)), with the added advantages of full control of orientation.

4 CONCLUSIONS

A data-driven approach to 3D visual speech synthesis based on captured 3D video of faces has been presented. Recent advances in 3D video capture have achieved simultaneous video-rate acquisition of facial shape and appearance. In this paper we have introduced face synthesis based on a graph representation of a phonetically segmented 3D video corpus. This approach is analogous to previous work in face synthesis by resampling 2D video (Bregler et al., 1997) and 2D video textures(Schödl et al., 2000). Face synthesis for novel speech utterances is achieved by optimisation of the path through the graph and concatenation of segments of the captured 3D video. A novel metric using a hierarchical wavelet decomposition is introduced to identify transitions between 3D video frames with similar facial shape, appearance and dynamics. This metric allows efficient computation of the similarity between 3D video frames for a large corpus to produce transitions without visual artifacts. Results are presented for facial synthesis from a corpus of 12minutes (18000 frames) of 3D video. Visual speech synthesis of novel sentences achieves a visual quality comparable to the captured 3D video allowing highly realistic synthesis without post-processing. The data-driven approach to 3D face synthesis requires minimal manual intervention between 3D video capture and facial animation from speech. Future extensions to the system introducing expression and secondary facial movements in a thoroughly engaging synthetic character are foreseen.

REFERENCES

Brand, M. (1999). Voice puppetry. In *Proceedings of SIGGRAPH '99*, pages 21–28, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Bregler, C., Covell, M., and Slaney, M. (1997). Video rewrite: driving visual speech with audio. In *Proceedings of SIGGRAPH '97*, pages 353–360, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Cao, Y., Faloutsos, P., and Pighin, F. (2003). Unsupervised learning for speech motion editing. In *Eurographics/ACM SIGGRAPH Symposium on Computer Animation '03*, pages 225–231.

Cohen, M. and Massaro, D. (1993). Modeling coarticulation in synthetic visual speech. In *Computer Animation '93*, pages 139–156.

Cosatto, E. and Graf, H. (2000). Photo-realistic talking heads from image samples. *IEEE Transactions on Multimedia*, 2(3):152–163.

Ezzat, T., Geiger, G., and Poggio, T. (2002). Trainable videorealistic speech animation. In *Proceedings of SIGGRAPH '02*, pages 388–398, New York, NY, USA. ACM Press.

Kalberer, G. and Van Gool, L. (2002). Realistic face animation for speech. *Journal of Visualization and Computer Animation*, 13(2):97–106.

Kovar, L., Gleicher, M., and Pighin, F. (2002). Motion graphs. In *Proceedings of SIGGRAPH '02*, pages 473–482, New York, NY, USA. ACM Press.

Kshirsagar, S. and Magnenat-Thalmann, N. (2003). Visyllable based speech animation. In *Eurographics'03*, pages 632–640.

Löfqvist, A. (1990). Speech as audible gestures. In Hardcastle, W. and Marchal, A., editors, *Speech Production and Speech Modeling*, pages 289–322. Kluwer.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, (264):746–748.

Schödl, A., Szeliski, R., Salesin, D. H., and Essa, I. (2000). Video textures. In *Proceedings of SIGGRAPH '00*, pages 489–498, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co.

Shashua, A. and Levin, A. (2001). Linear image coding for regression and classification using the tensor-rank principle. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*, pages 42–49.

Stollnitz, E., DeRose, T., and Salesin., D. (1995). Wavelets for computer graphics: A primer. 15:76–84.

Sumby, W. and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. 26:212–215.

Taylor, P., Black, A., and Caley, R. (1998). The architecture of the the festival speech synthesis system. In *Third International Workshop on Speech Synthesis*.

Wang, Y., Huang, X., Lee, C.-S., Zhang, S., Li, Z., Samaras, D., Metaxas, D., Elgammal, A., and Huang, P. (2004). High resolution acquisition, learning and transfer of dynamic 3-d facial expressions. In *Eurographics'04*, pages 677–686.

Ypsilos, I., Hilton, A., and Rowe, S. (2004). Video-rate capture of dynamic face shape and appearance. In 6^{th} *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 117–123.

Zhang, L., Snavely, N., Curless, B., and Seitz, S. (2004). Spacetime faces: high resolution capture for modeling and animation. *ACM Transactions on Graphics*, 23(3):548–558.