Talking Faces - Technologies and Applications

Joern Ostermann
University of Hannover
Hannover, Germany
ostermann@tnt.uni-hannover.de

Axel Weissenfeld
University of Hannover
Hannover, Germany
aweissen@tnt.uni-hannover.de

***Abstract*—This paper gives an overview of facial animation techniques. While facial animation is currently used in the entertainment and advertisement industry, it will become part of dialog systems in commercial applications. Animation techniques based on 3D models and image-based rendering are presented and evaluated by the following characteristics: automatism, realism and flexibility. While animation techniques based on 3D models provide for high flexibility and automatism, they often lack realism. Image-based techniques achieve photo realism, but lack in automatism and flexibility**.

Keywords—facial animation, overview, applications, 3d based models,image-based

I. INTRODUCTION

Computer aided modeling of human faces still requires a great deal of expertise and manual control to achieve realistic animations and to prevent unrealistic or non-face like results. Humans are very sensitive to any abnormal lineaments, so that facial animation remains a challenging task till this day. This paper gives an overview of the most important facial animation techniques, categorizes and compares them. One enhanced facial animation technique in each category is briefly discussed and evaluated. An evaluation has always to take into account potential applications of a particular animation technique. Hence, current and prospective applications and markets for facial animation are analyzed and explored. Especially prospective applications for facial animation combined with synthetic speech as part of modern dialog systems reveal great industrial opportunities.

Face animation research started in the early 70's by Parke [1]. In those days facial animation was limited by the available hardware, so that only primitive models describing the rough shape of a 3D face were animated. The first models were created by painting a set of polygons on a human face and taking 2D images from different views. A 3D-polygonal face (Fig. 1) is designed by reconstructing 3D points from corresponding 2D feature points. Hence, this algorithm requires manual assistance to generate 3D models and a patient human subject. Animation takes place by interpolating between two existing static expressions each defined by a simple 3D polygonal face. The animation described is very limited, since only intermediate expressions between known static expressions can be generated.

Figure 1: Two expressions of the same face, which are rendered using polygonal shading and Gourad's smooth shading algorithm [1].

In the last decade the quality of facial animation has significantly improved due to better computer systems, which have more power and increasing software capabilities. The animation techniques range from animating 3D models to image-based rendering of models. The techniques are so diverse because of the very different commercial applications. Furthermore, hardware components vary. For instance, hardware components for creating facial models range from 3D laser scanners to a single or multi camera set-up. Nowadays photo realistic automatic facial animation is possible. We define photo realism as the goal of synthesizing facial animations, which are undistinguishable from real photographs or videos.

In the remainder of this paper, we analyze current and prospective applications and markets for facial animations (Section 2). We then discuss different facial animation techniques based on 3D models (Section 3). In Section 4 different image-based facial animation systems are presented. The different techniques described in Sections 3 and 4 are characterized and compared in Section 5.

II. APPLICATIONS AND MARKETS

Today computer animation is mostly used in the entertainment and media industry, which will reach revenue of $1.4 trillion in 2007 from $1.1 trillion in 2002, a 4.8% percent average annual growth rate [6]. Facial animation entertains humans in various commercial applications, such as virtual characters in computer games and movies. These animation techniques need to be flexible to satisfy the requirements of the entertainment industry. For example, animation techniques should be able to animate 'human-

like' and 'cartoon-like' characters and render them from arbitrary views. For pure entertainment, humans usually expect realistic but not photo realistic facial animations, which are available nowadays.

In the last years facial animation techniques made great progress that will open new markets for face animations. Most techniques have focused on improving the realism of lip movements, since the mouth area is very important for realistic animations. Synchronizing lip movements with spoken output is a challenging task, since coarticulation effects describing the influence of consecutive phonemes onto each other do not allow a simple mapping of phonemes to visemes. Visemes are mouth shapes that correspond to the phonemes of the spoken words. Nevertheless some modern facial animation techniques achieve photo realism. Furthermore, recent advances in speech recognition, natural language understanding and speech synthesis give the opportunity to use talking heads, which combine text-to-speech synthesis (TTS) with facial animation, as part of a modern human-machine interface in modern dialog systems (Fig. 2). The trust and attention of humans toward machines increases if humans are communicating with talking heads instead of text-only output by 30% [3] [13]. Photo realistic facial animations offer new commercial applications as part of modern dialog systems. For instance, talking heads can be used as a newsreader, or a virtual assistant as part of an e-commerce, e-learning or e-care (costumer management relationship) web site. These prospective markets will grow continuously. We expect that the facial animation industry will also participate in this growth. For instance, the e-care industry is focusing on self-service solutions for customers, which can be realized by knowledge management (KM) systems or customer relationship management (CRM) systems. As a result companies are investing money to build KM systems, which will carry the worldwide market of KM software from $1.4billion in 1999 to $5.4billion in 2004 (source: Frost&Sullivan). Facial animation as part of dialog systems will help enable self-service capabilities and participate in this growth.

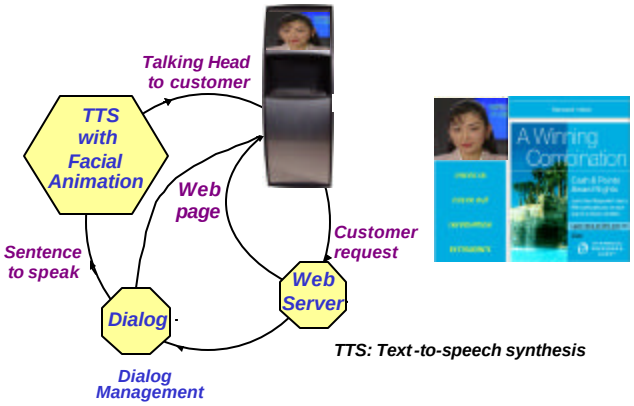

Figure 2: An internet-based customer service as a modern interactive service, which returns a web site with an integrated talking head [3].

III. FACIAL ANIMATION BASED ON 3D-MODELS

Nowadays a great variety of different animation techniques based on 3D models exist. In general, these techniques first generate a 3D model consisting of a 3D mesh, which defines the geometric shape of a face. For that a lot of different hardware systems are available, which range from 3D laser scanners to multi camera systems. In a second step, either a human-like or carton-like texture may be mapped onto the 3D mesh. Besides generating a 3D model, animation parameters have to be determined for the later animation. Animators may also use professional computer animation software to design and animate 3D models.

MPEG-4 is the first international standard that supports facial animation. It describes the definition, encoding, transmission and animation of 3D face models. Furthermore, text-to-speech (TTS) synthesizers can be integrated into the facial animation part of an MPEG-4 decoder for interactive services [14].

Facial animation techniques based on 3D models can be clustered into 3 major groups: morphing static 3D models, performance-based animations and physics-based models. Each major group will be described and one of the most sophisticated animation techniques in each group will be briefly explained. However, many modern animation systems merge different techniques in order to improve facial animation.

III.1 Morphing static 3D models

Animation techniques based on static 3D models were already developed in the 70's [1][15]. New facial expressions may be synthesized by interpolating between two static facial expressions. The quality of these animation techniques has significantly improved especially through more sophisticated 3D models and new texture-mapping methods in recent years [28][29][30]. The combination of different static facial expressions looks surprisingly realistic nowadays, whereas a realistic speech animation is not possible yet.

The technique described by Pighin et al. [29] synthesizes facial expressions from photographs. First a static 3D model is generated for each facial expression. Several cameras capture the human subject showing different specified facial expressions such as joy or sadness from different views. A generic head model is fitted to an individual human head in three steps. First the extrinsic camera parameters and focal length are calculated for each view. Furthermore, the mesh vertices of the generic head model are fitted to 13 hand-marked feature points in the images, for which the 3D coordinates are reconstructed. Once the 13 vertices of the generic head model are fitted, the other vertices of the head are deformed to fit. In a last step the shape of the face model is refined in more detail by selecting additional correspondences in the 2D images. The texture-maps from the different images for the 3D model are extracted and may be used for a view-dependent or view-independent texture-mapping. The view-dependent texture-mapping method

gives the opportunity to render realistic 3D models by taking into account the textures of two views (or more) enclosing the rendered view. The textures are weighted by observing the angle of the rendered view and the enclosing two views. Other facial parts such as eyes, ears and hair are textured in a separate process, since for instance the eyes and teeth are usually partially occluded by the facial tissue. For animation this technique generates continuous transitions between facial expressions by morphing between corresponding head models. The morphing between arbitrary static 3D models is simplified, since the topology of the face meshes is identical for all static 3D models. Hence, a 3D morphing is performed by a linear interpolation between the geometric coordinates of corresponding vertices in each of the two face meshes. Then the face model obtained by geometric interpolation is independently rendered with the first and second texture and the resulting images are blended together. Furthermore, the technique enables the user to combine facial features from different facial expressions (Fig. 3).

Figure 3: This figure shows the different steps necessary to design a "debauched smile" expression. The left side shows the facial parts of the original expressions (highlighted), which are attached to the neutral expressions on the right side. The different stages of the design are shown on the right side [29].

III.2 Performance-based animations

These animation techniques capture the three-dimensional geometry and color information of human facial expressions. Furthermore, motion information is gathered by measuring real human motions to drive synthetic characters. For that a laser- or video-based motion-tracking system such as a multi-camera set-up [7] is commonly used, in which a human's face is marked by several feature points, which are tracked over an image sequence. The captured motion, 3D geometry of the head and color information are used to synthesize realistic facial animations. Early works in this area are discussed in [16][17][18]. More advanced methods are described in [2][19][20].

The animation technique developed by Kalberer et al. [2] focuses on realistic face animation for speech and provides great flexibility to the user. The animation pipeline provides 3D face models for a number of individuals differing in age, race and gender, which are captured by a structured light system. After the generic face model is fitted to scanned 3D face models, a principal component analysis (PCA) is performed to find the principal components of each individual head model. The obtained components span a space, called face space, in which the average face is located at the origin. The user of the system can choose between fitting the generic head model to a scanned 3D face model of a human subject and designing a new face through the combination of principal components in face space. The speech dynamics are measured by tracking 116 feature points of the face in 3D. In this way the exact 3D motion of facial features while speaking can be investigated. The mouth shape of each frame can be identified as a single viseme. Each viseme is processed by an independent component analysis and characterized by its independent components, which span a viseme space. For ten reference faces the speech dynamics and their visemes, which are represented by a point in the viseme space, are already available to the user. Besides representing ten points in face space, these reference faces span a hyper-plane in face space. Faces for which no 3D dynamics are measured can be animated by linearly approximating the face through the ten given reference faces, that provide the speech dynamics necessary for an animation. A face is animated by determining the appropriate trajectories for given phonemes. The trajectories lead from one viseme to the next with the right timing in viseme space. The animation of the mouth is improved by considering the biomechanics of the articulation of the chin, which is a physics-based animation (section III.3). This animation technique includes natural head motion captured from reference faces and the user may add basic emotions afterwards.

Figure 4: Synthesized sample of facial animation [2].

III.3 Physics-based animation

In this approach the 3D model has an underlying anatomical structure. Different layers of muscles, cartilage, bones, nerves, blood vessels, glands, fatty tissue, connective tissue, hair and skin describe the anatomy of the head, such that a model based on the anatomy allows a deformation of the head in anthropometrically meaningful ways. However, until this day no facial animation technique based on the described level of anatomical detail has been developed. As a result 3D models based on the anatomy simplify the anatomical structure of the head resulting in a lower animation quality. The physics-based approach goes back to the 80's [21][22][23][24] and is still used and improved in recent animation techniques [25][26][27].

The animation technique described by Kaehler et al. [25] is a sophisticated physics-based technique. A generic head model is defined consisting of five major structural components: a triangle mesh for the surface, a layer of virtual muscles to control the animation, an embedded skull including a movable lower jaw, to which skin and muscles attach, a mass-spring system connecting skin, muscles and skull, and separately modeled components like eyes, teeth and tongue. The generic head model has to be fitted to the individual head, which may be obtained by a range scanning device. Therefore the skin, skull and muscles of the generic head model are deformed to the target face model. Muscle contraction parameters are specified for some representative facial expressions. Facial motion is mainly controlled by specifying muscle contraction over time. Furthermore, this approach allows generating speech-synchronized contraction parameters of the jaw, tongue and mouth for a reference head model [31]. These parameters define the muscle contractions of each phoneme resulting in a corresponding viseme. The parameters defined for the reference face can simply be reused by animating new models.

Figure 5: Head models with anatomical structure [25].

IV. IMAGE-BASED FACIAL ANIMATION

Image-based rendering processes only 2D images, so that animations are synthesized by combining different facial parts of recorded 2D images. Hence, a 3D model is not necessary for animations. In general, image-based facial animations consist of two main steps (Fig. 7): Audiovisual analysis of a recorded human subject and synthesis of facial animation. In the analysis step a database with images of deformable facial parts of the human subject is collected, while the time-aligned audio file is segmented into phonemes. In the second step a face is synthesized by first generating the audio from the text using a TTS synthesizer. The TTS synthesizer sends phonemes and their timing to the face animation engine, which overlays facial parts corresponding to the generated speech over a background video sequence. Background sequences are recorded video sequences of the human subject with typical short head movements.

The first approaches of image-based facial animation started at the beginning of the nineties [4][5]. In this chapter three sophisticated image-based facial animation techniques are briefly introduced. The results for some image-based facial animations are shown in Fig. 6 [12].

Figure 6: Image-based face animation [12].

The animation technique developed by Bregler et al. [8] consists of two steps: an analysis and a synthesis stage. In the analysis step a database of short video clips, each showing three consecutive phonemes (called triphone), is generated. All triphone videos showing the mouth and chin areas are labeled with their phonemes, which are determined by aligning phonemes to the video, and fiduciary-point locations around the mouth and jaw. Before fiduciary-point locations are calculated, each face image is warped to a standard reference frame in order to compensate for head motion during recording. The fiduciary-point locations are determined by eigenpoint models, which are partly manually and partly automatically generated. An eigenpoint model consists of 54 eigenpoints telling the position of the mouth and jaw. Altogether over 300 eigenpoint models are generated during a training session, which are then used for the analysis of a recorded video sequence in order to determine the position of mouth and jaw. The triphone videos in the database are labeled with the phoneme sequence and the locations of fiduciary points.

In the synthesis step a new video is synthesized by labeling a new speech track and selecting the most appropriate triphone videos. Triphones are selected by minimizing an error function, which takes into account the phoneme-context distance and the distance between lip shapes in overlapping visual triphones. Hence, this error function verifies that phonemes from the same context with similar mouth shape are selected for the synthesized video. The triphone videos are aligned with the phoneme script from the audio file in order to synchronize lip movement

with spoken output. In a last step the triphone videos are placed into a background sequence. For that the triphone videos are warped to the background and lip shapes of overlapping triphones are cross-faded.

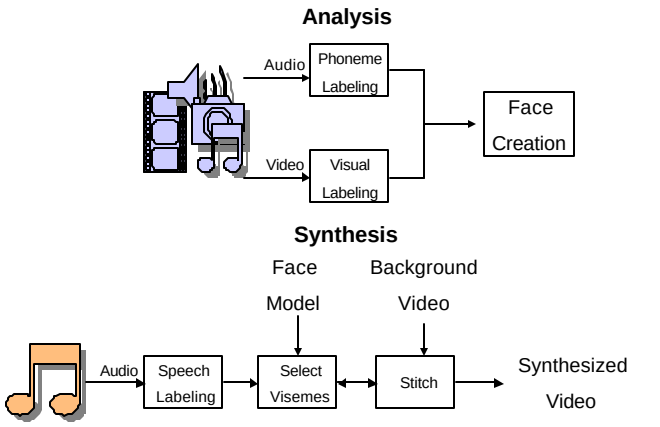

Figure 7: Overview of analysis and synthesis stage for sample-based face animation.

A second image-based technique was developed by Ezzat et al. [9][10]. This facial animation system has an analysis and synthesis step, each consisting of three minor processing steps, as shown in Fig. 8. The analysis starts with preprocessing the recorded video data (corpus) by aligning phonemes of the audio file and normalizing the images of the videos such that the face appears in a normalized position and orientation. The heart of this animation technique is a multidimensional morphable model (MMM) representation, which is capable of morphing between various basic mouth shapes. The MMM is automatically built by performing a principal component analysis (PCA) on all recorded and normalized images and clustering the obtained PCA parameters. 46 prototype images are automatically selected to represent the corpus of the recorded subject. However, these prototype images do not have an explicit relationship to visemes, and instead give a basic set of image textures. Furthermore, 46 optical flow correspondences are computed, which represent the correspondences between one reference prototype image and the others in the MMM. Each phoneme is represented by a region within the MMM space with a particular position and covariance. The covariance takes into account observed coarticulation effects in the corpus and pulls the mouth trajectories to the best phonetic regions. The mouth trajectories describe the motion, smoothness, dynamics and coarticulation effects of the mouth animation. The entire recorded corpus, altogether s images, is mapped onto the constructed MMM space during the analysis. A time series of $(a_j, b_j)_{j=1}^{s}$ parameters, in which a and b are two 46-dimensional parameter vectors identifying the mouth shape and mouth texture respectively, describe the trajectories of the original lip motion in the MMM space.

The synthesis step starts with mapping an input phoneme stream to the mouth trajectory (a, b)-parameters in the MMM space, while minimizing a function consisting of the target term and smoothness. Then the MMM is used to synthesize the new visual output from the trajectory parameters. First a synthetic optical flow correspondence consisting of linear combinations of the prototype flows is warped forward. Then the prototype images are warped forward by considering the synthetic correspondence. Finally the warped prototype images are morphed to a final image. After all mouth shapes for the input phonemes are generated, the novel mouth shapes are placed onto a background sequence.

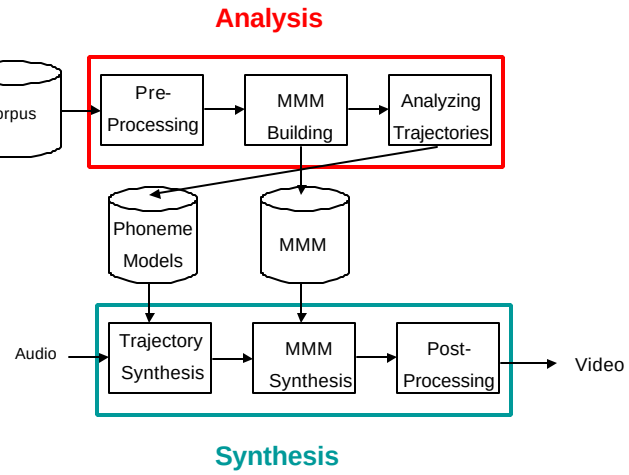

Figure 8: Overview of analysis and synthesis stage [10].

Another image-based facial animation technique was developed by Cosatto et al. [11] [12] [32]. This technique generates synthetic images by texture-mapping different facial parts modeled in 3D to a background sequence modeled by a single plane. Thus, this technique is more flexible than the previous two techniques while preserving photo realism. Sometimes this technique is also notated as a 2.5D technique referring to the combination of 2D and 3D animation methods.

For face model creation a talent is recorded in a studio. Moreover, a basic 3D mask describing the shape of the talent's head has to be generated, e.g. by a 3D laser scanner. The analysis of the recorded sequence starts by aligning phonemes to the recorded video. Through image analysis significant facial features like mouth and eyes are located and extracted. These facial parts have to be normalized, before they are stored in a database. Normalization means to compensate head motion that is estimated by tracking the aligned 3D mask along the video sequence. Before the extracted facial features are stored in a database, each feature is labeled by its geometric features, PCA features and phonetic information. Each image of the recorded video is labeled with the same information as well as with the head pose of the face in the image.

The synthesis starts with a text-to-speech system synthesizing spoken output and a target phoneme string. Then an animation graph is built that has the appropriate facial features for the target phoneme string. A Viterbi search algorithm is selected to find the best mouth samples

in the database that provides the lowest target costs. Target costs define the transition costs between two mouth samples taking into account the phonetic context and visual difference between two samples. Furthermore, this technique allows for controlling visual prosody, by animating the head motion, eye blinks and inserting expressions. The best mouth samples and other different facial parts are mapped onto the primitive 3D mask (Fig. 9), which matches the pose exhibited by the head in the background sequence.

Figure 9: The side views hints at the assembly process of 3D parts and background sequence [12].

V. Evaluating Facial Animation Techniques

First the two approaches for facial animation introduced in Sections 3 and 4 are evaluated: facial animation based on 3D models and image-based facial animation. Both approaches have common characteristics as automatism, flexibility and realism. Automatism evaluates the process of model creation and the synthesis of animation. Flexibility also includes different characteristics, such as synthesis speed or rendering the face model from arbitrary views and with user-defined textures. Realism evaluates how closely an animated face matches a real face. It can be further divided into the realism of the head, lip movements and variety of facial expressions.

The strengths and weaknesses of facial animation techniques based on 3D models and image-based rendering are presented in Table 1. Animation techniques based on 3D models have in general a much higher automatism than image-based models. That is because acquiring and fitting a 3D model from new human subjects to the generic head model is highly automatic. Alternatively professional computer animation software may be used to quickly design 3D models for animation. Image-based techniques require a tedious recording of each human subject in a professional studio environment and sophisticated analysis of the recorded video.

Techniques based on 3D models have a great flexibility, since 3D models may be rendered from arbitrary views and with user-defined textures. Image-based facial animations do not offer this flexibility, because animations can only be synthesized from existing data samples. Therefore the user cannot arbitrarily choose the texture and view.

Table 1: Comparison of facial animations based on 3D models and image-based models.

Characteristics	3D models	Image-based
Automatism	high	low
Flexibility	high	low
Realism	medium	high

However, techniques based on 3D models have in general a lack of realism; especially mouth animations do not look realistic. As long as the head only fulfills rigid movements a high quality animation may be obtained. However, when plastic deformations occur, the realism decreases very quickly. The meshes of the 3D model are changed, while the texture-maps of the 2D images remain the same, so that artifacts occur during the rendering. While surfaces with little texture may only show tolerable deformations, the animation of the human mouth causes high deformations and the animation may look artificial. That is because the mouth area shows a lot of small wrinkles during lip movements, which cannot be properly modeled yet. As a result the focus of many current animation techniques is to improve the animation of the mouth, so that realistic lip synchronization to spoken output is generated. Image-based facial animation techniques achieve great realism in synthesized videos. In particular some image-based speech animation cannot be distinguished from recorded videos [32].

Animation techniques based on 3D models are mainly used in the entertainment industry, in which a photo realistic animation is not necessarily required but flexibility. Image-based facial animation techniques achieve photo realism and therefore may find commercial applications as part of modern dialog systems.

The facial animation techniques based on 3D models discussed in Section 3 and image-based facial animation techniques described in Section 4 are benchmarked in Tables 2 and 3, respectively, to specify in more detail advantages and disadvantages of each technique.

Pighin et al.'s animation technique focuses on combining different static facial expressions. The animation technique synthesizes realistic facial expressions using view-dependent texture-mapping. However, generating the 3D model may need the user intervention. Moreover, realistic speech animation is not possible, since the correct texture cannot be provided for all typical mouth shapes.

The animation technique developed by Kalberer focuses on speech animation and achieves the highest realism, if only animation techniques based on 3D models are compared. As soon as the face and viseme space are generated the creation of new models is highly automatic.

The animation system also includes other techniques, e.g. the motion of the lower jaw is physics-based controlled.

Table 2: The three facial animation techniques based on 3D models described in Section 3 are compared. The characteristics are grouped in automatism [A], realism [R] and flexibility [F].

Characteristics	**Pighin [29]**	**Kalberer [2]**	**Kaehler [25]**
Model creation, synthesis [A]	medium	high	medium
Synthesis speed [F]	medium	medium	medium
Arbitrary view, texture [F]	high	high	high
Head [R]	high	medium	medium
Lip animation [R]	low	medium	low
Variety of expressions [R]	high	high	high

Kaehler et al.'s physics-based animation technique deforms the head by muscle contraction parameters. The contraction parameters are determined once for a reference face and afterwards all other 3D models can be animated by using these parameters, such that the synthesis is highly automatic. Generated facial expressions look quite realistic, while mouth animations still look unnatural.

In the field of image-based facial animation, Bregler et al.'s synthesized videos look very realistic with appropriate lip movements. A disadvantage of his technique is the large database necessary for storing mouth samples (triphones).

Ezzat et al.'s technique synthesizes very realistic facial animations with a very small database, which increases the flexibility of this technique. However, the synthesis speed is low. Furthermore, this technique cannot generate any facial expressions.

The animation technique developed by Cosatto et al. enables photo realistic animations. The user can select different facial expressions and control facial motions like head motions and eye globe motions. The synthesis of facial animation is possible in real-time, which is an important criterion for applications in dialog systems. However, a large database storing the facial features is required.

The latest facial animation systems show the phenomenon, that more and more different techniques are combined. In this way the advantages of each technique are fully used and the animation is improved. For instance, Cosatto et al. combine a 3D model with image-based rendering. In this way, they achieve photo realistic facial animations while increasing the flexibility.

Table 3: Image-based facial animation techniques introduced in Section 4 are compared. The characteristics are grouped in automatism [A], realism [R] and flexibility [F].

Characteristics	**Bregler [8]**	**Ezzat [10]**	**Cosatto [12]**
Model creation, synthesis [A]	low	low	low
Synthesis speed [F]	high	low	high
Arbitrary view, texture [F]	low	low	medium
Head [R]	high	high	high
Lip animation [R]	medium	medium	high
Variety of expressions [R]	low	low	high

VI. CONCLUSIONS

In this paper current and prospective applications for facial animations are presented. Animations are mainly used in the media and entertainment industry nowadays, but more sophisticated animation techniques lead to prospective applications, e.g. in modern dialog systems. Facial animation techniques based on 3D model and image-based techniques were discussed. Each approach has obviously advantages and disadvantages. Techniques based on 3D models impress by their great automatism and flexibility while lacking in realism. Image-based facial animation achieves photo realism while having little flexibility and lower automatism. The image-based techniques seem to be the best candidates for leading facial animation to new applications, since these techniques achieve photo realism. The image-based technique combined with a 3D model generates photo realistic facial animations, while providing some flexibility to the user.

REFERENCES

[1] F.I. Parke, “Computer generated animation of faces”, in ACM National Conference. ACM, November 1972.

[2] G. Kalberer, "Realistic Face Animation for Speech ", PhD Thesis, Swiss Federal Institute of Technology Zurich, 2003.

[3] J. Ostermann, A. Weissenfeld, “Face Animation for Human Computer Interfaces”, Proceedings of WIAMIS 2004, 2004.

[4] D. Beymer, A. Shashua, T. Poggio, “Example based image analysis and synthesis”, A.I. Memo No. 1431, Artificial Intelligence Laboratory, MIT, 1993.

[5] K. Scott, D. Kagels, S. Watson, H. Rom, J. Wright, M. Lee, and K. Hussey, “Synthesis of speaker facial movement to match selected speech sequences”, In Proceedings of the Fifth Australian Conference on Speech Science and Technology, volume 2, pp. 620-625, 1994.

[6] PricewaterhouseCoopers, “Entertainment and Media Outlook: 2003-2007, Global Overview and North America”, edition.

[7] Eyetronics. Http://www.eyetronics.com, 2003.

[8] C. Bregler, M. Covell, M. Slaney, “Video Rewrite: Driving Visual Speech with Audio”, Proc. ACM SIGGRAPH 97, in Computer Graphics Proceedings, Annual Conference Series, 1997.

[9] T. Ezzat, T. Poggio, “MikeTalk: A Talking Facial Display Based On Morphing Visemes”, Proc. IEEE Computer Animation, pp. 96-102, 1998.

[10] T. Ezzat, G. Geiger, T. Poggio, “Trainable Videorealistic Speech Animation”, Proc. ACM SIGGRAPH, pp. 388-397, 2002.

[11] E. Cosatto, H.P. Graf, “Sample-Based Synthesis of Photo-Realistic Talking heads,” Proc. IEEE Computer Animation, pp. 103-110,1998.

[12] E. Cosatto, H.P. Graf, “Photo-realistic talking heads from image samples”, IEEE Trans. on Multimedia, vol. 2, no. 3, pp. 152-163, Sept. 2000.

[13] Ostermann, D. Millen, “Talking heads and synthetic speech: An architecture for supporting electronic commerce”, Proc. ICME, pp. MA2.3, 2000.

[14] J. Ostermann and M. Beutnagel, A. Fischer, Y. Wang, "Integration of talking heads and text-to-speech synthesizers for visual TTS", ICSLP 99, Australia, December 99.

[15] F.I. Parke, “A Parametric Model for Human Faces”, Ph.D. thesis, University of Utah, Salt Lake City, UT, December 1974.

[16] B. deGraf, “Notes on facial animation”, In State of the Art in Facial Animation, SIGGRAPH’89 Tutorials, Volume 22, pp. 10-11. ACM, NewYork, 1989.

[17] P. Bergeron, P. Lachapelle, “Controlling facial expressions and body movements”, In Advanced Computer Animation, SIGGRAPH’85 Tutorials, Volume 2, pp. 61-79. ACM, NewYork, 1985.

[18] H.D. Kochanek, R.H. Bartels, “Interpolating splines with local tension, continuity and bias control”, Computer Graphics, 3(18): 33-41, 1984.

[19] B. Guenther, C. Grimm, D. Wood, H. Malvar, and F. Pighin, “Making Faces”, In Proc. SIGGRAPH, pp. 55-66, 1998.

[20] I. Lin, J. Yeh, M. Ouhyoung “Realistic 3D facial animation parameters from mirror-reflected multi-view video”, In Proc. Computer Animation 2001 Conf., pp. 2-11, 2001.

[21] S.M. Platt, N. I. Badler, “Animating facial expressions”, Computer Graphics, 15(3):245-252, 1981.

[22] K. Waters, “A muscle model for animating three-dimensional facial expressions”, Computer Graphics (SIGGRAPH’87), 21(4):17-24, July 1987.

[23] S.D. Pieper, “More than skin deep: Physical modeling of facial tissue”, Master’s thesis, Massachusetts Institute of Technology, Media Arts and Sciences, Cambridge, MA, 1989.

[24] D. Terzopoulos, K. Waters, “Physically-based facial modeling analysis and animation”, J. of Visualization and Computer Animation, 1(4):73-80, March 1990.

[25] K. Kaehler, J. Haber, H. Yamauchi, HP. Seidel, “Head Shop: Generating animated head models with anatomical structure”, In Proc. 2002 ACM SIGGRAPH, Symposium on Computer Animation, pp. 55-63, 2002.

[26] Y. Lee, D. Terzopoulos, K. Waters, “Realistic Modeling for Facial Animations”, Computer Graphics (SIGGRAPH’95 Conf. Proc.), pp. 55-62, 1995.

[27] K. Waters, J Frisbie, “A coordinated muscle model for speech animation”, In Graphics Interface, 1995.

[28] D. Chen, A. State, D. Banks, “Interactive shape metamorphosis”, In Symposium on Interactive 3D Graphics, editor, SIGGRAPH’95 Conference Proceedings, pp. 43-44, 1995.

[29] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D. H. Salesin, “Synthesizing realistic facial expressions from photographs”, In SIGGRAPH’98 Proceedings, pp. 75-84, 1998.

[30] V. Blanz, T. Vetter, “A morphable model for the synthesis of 3d faces”, In Proc. SIGGRAPH, pp. 187-194, 1999.

[31] I. Albrecht, J. Haber, K. Kaehler, M. Schroeder, H.-P. Seidel, “may i talk to you? :-)" - facial animation from text”, Proceedings of the 10th Pacific Conference on Computer Graphics and Applications (Pacific Graphics 2002), Tsinghua University, Beijing, 2002.

[32] E. Cosatto, J. Ostermann, H. P. Graf, J. Schroeter, “Lifelike Talking Faces for Interactive Services,” Invited Paper, Proc. of the IEEE, Special Issue on Human-Computer Multimodal Interface, Vol. 91, No. 9, pp. 1406-1429, Sept. 2003.