Talking Head: Synthetic Video Facial Animation in MPEG-4.

A. Fedorov, T. Firsova, V. Kuriakin, E. Martinova, K. Rodyushkin and V. Zhislina
Intel Russian Research Center, Nizhni Novgorod, Russia

Abstract

We present a system for facial modeling and animation that aims at the generation of photo-realistic models and performance driven animation. It is practical implementation of MPEG-4 compliant Synthetic Video Facial Animation pipeline (Simple and Calibration Profiles with some modifications), which includes: facial features recognition & tracking on real video sequence; obtaining, encoding, network transfer and decoding of the facial animation parameters (FAPs); face model adaptation (calibration) and texturing by photos or video stream; synthetic 3D head animation and visualization.
All pipeline stages, except model adaptation, operate in real time, fully automatically and produce appropriate results.
To obtain the animation parameters new methods of robust online face tracking and more precise but off-line FAPs estimation are developed.
Model-independent animation method was used. A number of new features including mouth contours special processing, eyelashes and "expression wrinkles" were added beyond standard to improve the resemblance of the Talking Head to the prototype person.

The following libraries and applications were created to encapsulate "Talking Heads" original technologies: Face Tracking&Analysis Library, Images-to-Head Calibration application, Intel Facial Animation&Visualization Library. They are absolutely independent and could be used separately for various applications.

Keywords: MPEG-4, FAPs, Facial Animation

1. INTRODUCTION

MPEG-4 is the modern coding standard [1] [2], which supports among others special synthetic objects such as 3D human heads, animated by stream of Facial Animation Parameters (FAPs).
There are many reasons why these MPEG-4 synthetic video Talking Heads are sure to find a paying place in the market. Some of them are:

- Extremely low data traffic (500 times better than MPEG-1), which is critical for low bit rate channels,
- New opportunities limited only by imagination, flexibility superior to ordinary real video.

These advantages shape possible fields of application of the Talking Heads, - advanced telecommunication (3G phones and video-conferencing), computer games, filmmaking, e-services, narration and education.
A lot of companies and research centers work over Talking Heads related problems, for example [3]. Some commercial products offer the possibility of personalized Talking Heads creation. However as far as we know the technology of automatic creation and real-time rendering of Talking Heads with photo realistic appearance and animation doesn't exist. The majority of applications provide good quality of their models using manual (artist) work for every model. One more widespread approach is to create models automatically, but using frontal image only based technique that is much more simple but unusable if we want to get model side view. Moreover many applications have a small window size and don't permit any model zoom to evaluate synthetic video quality. When talking about animation it's necessary to notice many applications use either manual contribution in animation process (target morphs manual creation or muscles contraction correction) or bound model animation by Lip Synch technology only making all models talk the same way without individual features. All mentioned above restrictions significantly narrows the field of these applications employment. And the primary goal of our project was to obtain high quality synthetic video i.e. to mimic real person as much as possible in appearance and facial motion.
We developed facial animation pipeline mainly conformant to the MPEG-4 Calibration profile [2]. The all stages of our pipeline, except model adaptation and animation table calculation, operate in real time, fully automatically and work for arbitrary model.
One of the main restrictions of MPEG-4 facial animation is its definition via feature point's displacements only, while in real life human face, driven by muscles displays a great variety of expressions which is difficult or even impossible to describe by feature point's displacements only. E.g. feature point's displacements don't allow explicit specification of subtle but important for perception effects such as skin expression wrinkles. Moreover, some feature points (e.g. on cheeks and eyelids centers) are ambiguously determined in MPEG-4 standard specifications and therefore are hard to interpret by algorithmic realization. So, they are very difficult to recognize and track automatically that leads to wrong estimation of their movements.
We employ another approach to MPEG-4 compatible model adaptation, animation and visualization – to use feature contours – contours passing through feature points to enhance model adaptation, animation and visualization. Generally such approach is more robust since the exploitable information is not local (points) but more global (contours).
In model adaptation process contours allow more natural representation of the topological structure of a head and therefore more precise modeling of prototype person. Using of contours in our model adaptation approach is described in [4],[5].

Note, that the MPEG-4 standard defines FAP set, and methods of its compression/decompression only. Generation of FAP stream from captured images, model calibration and animation/visualization are outside of the standard. We accent namely these fields, and present specific methods and algorithms of face localization, recognition and tracking of the FPs automatically in the input video sequence, as well as the methods of geometry calibration, texturing, animation and visualization.

2. OUR TECHNOLOGY IMPLEMENTATION

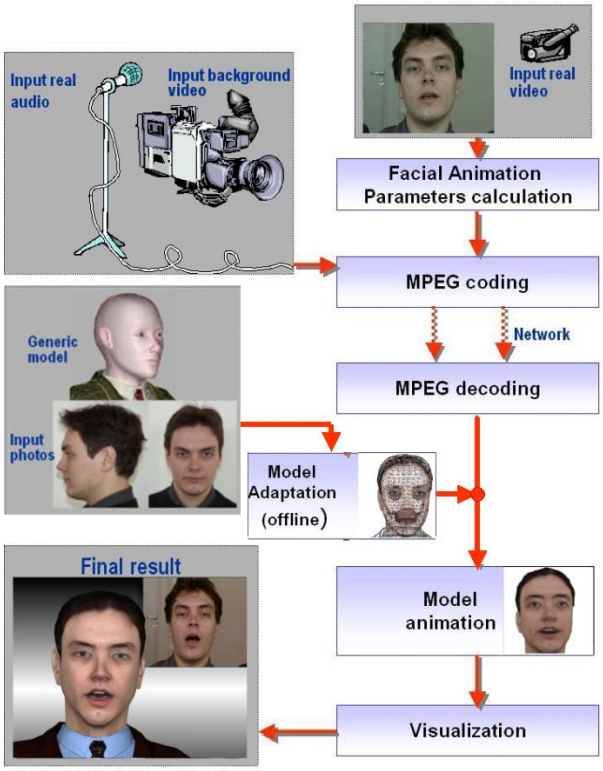

Figure 1. Talking Head Technology Pipeline

The scheme of our pipeline is presented on **Figure 1**.

- The first stage of the pipeline, which takes place in the encoder, is FAPs estimation from input video sequence. Note, that the MPEG-4 standard does not define methods for FAP stream generation. So, in order to generate a FAP stream, we have developed new and used existing methods and algorithms for automatic face localization, tracking and FAPs estimation from the input video sequence. This is an understandably challenging problem, considering the precision and accurateness required for correct FAPs estimation for a specific animation task. Algorithms used for this stage are contained in Face Tracking and Analysis library (FaceTA), which provides tools for fully automatic face location, tracking and FAPs estimation.
- The next stage is the compression and transfer of the resulting stream to the decoder in full conformity with MPEG-4 specifications.
- Then in the decoder the FAP stream is decompressed, that is a proper FAP values are obtained for each frame to run the 3D model animation. Since the MPEG-4 standard does not define a method for face animation using FAPs we have developed algorithms for FAPs interpretation and special processing methods to make facial animation more realistic.
- The final stage is rendering by means of OpenGL standard library. Developed algorithms for model animation and rendering are contained in 3D Face Animation & Visualization library (IFAL).
- Also the offline stage of generic face model calibration to prototype person (including texture generation) was implemented. Calibration takes place on client side –in FAPs decoder. Personalized models can be created from two photos taken in front and profile directions, or photos taken in arbitrary directions or a video showing head rotation from one profile to the other. Current Implementation requires certain user assistance in selecting feature elements on images – the rest of the process is performed automatically.
- Developed technologies have been implemented in Head Calibration Environment (HCE) and Video-to-Head (V2H) Environment providing user interfaces to them.

The following libraries and applications encapsulate Talking Head original technologies and are ready for external customers delivery

- Face Tracking&Analysis Library (FaceTA)
  Library encapsulates automatic human face recognition & tracking functionality.
- Head Calibration environment (HCE) and Video-to-Head environment (V2H)
  Interactive applications for generic model calibration (personalization) by prototype person's photos and video stream.
- Intel Facial Animation Library (IFAL)
  Library encapsulates 3D model transformations, model independent animation & visualization functionality.

Figure 2 shows FaceTA & IFAL libraries and HCE application places in our pipeline.

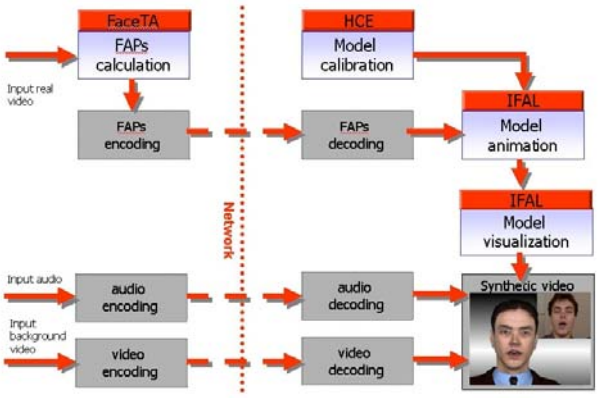

Figure 2. Libraries usage

Note that none of the libraries & applications presented support the following functionality:

- FAP stream coding \ decoding according to MPEG standard. FaceTA generates and IFAL operates with FAP values merely. To code \ decode FAP stream the special additional library is required (eg Intel Media Processing Library (MPL))
- Any activity concerned with audio (playback, FAP extraction etc)
- Any synchronization tasks (FAP and real video stream etc)

Two main types of applications were developed to demonstrate the results of our pipeline:

- MPEG4 “Talking Head” Player (to playback pre-recorded FAP sequences)
- full-duplex synthetic video-conferencing (to provide real-time communication via synthetic talking heads).

In the input these applications need VRML models file, FAPs file (in case of pre-recorded file) or stream (videoconferencing), natural audio and video MPEG-coded files, while the output is a synthetic video sequence showing a “talking head”, synchronized with audio stream drown over static image or a real video sequence background. Model animation is synchronized with audio and video streams according to MPEG standards.

3. HEAD MODEL

According to MPEG-4 standard specifications, a human head is a synthetic visual object whose representation is based on VRML standard [6]. Each decoder under MPEG-4 has its own face model called “generic model”. The generic model in our case (Figure 3) was selected from the “Planet People” CD [7] and modified to conformity with MPEG-4. Currently the Face Model Scene Graph includes a group of standard-conforming eight objects (skin, eyes, pupils, teeth, tongue) and some additional objects for enhanced realism (hair, glasses, mouth cavity and shoulders): up to 8,000 vertices and 16,000 triangles in total.

In MPEG-4 Calibration profile generic model adaptation to the prototype person is required. This process is called calibration and is not specified in MPEG-4. In our case the initial inputs for adaptation are the generic model and few images of the prototype or video stream taken with a digital video camera. All model objects are texturized with the total texture map volume of about 4Mb.

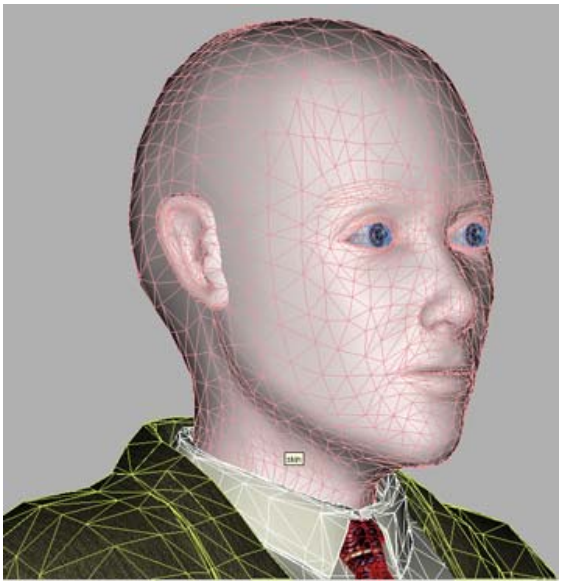

Figure 3. Face Model

4. MODEL ADAPTATION PIPELINE

The calibration process is designed as a pipeline of sequential stages that register available views in a framework associated with the head, adjust the geometry of a model in order to obtain precise matching with the input data, and generate a consistent texture by merging the texture candidates obtained through inverse texture mapping.

The input data of the calibration process can be either

- A pair of images taken at front and profile directions, or
- A collection of images taken at arbitrary directions, or
- A video stream having front and both profile views.

In [4] - model calibration pipeline description in detail.

5. FACE DETECTION, TRACKING AND FAP ESTIMATION

Special attention in our system was paid to algorithms for automatic localization and tracking of the face and its elements in the input video sequence in order to generate a FAP-stream.

For automatically FAPs estimation from video sequence the existing methods and new original algorithms are used conjointly.

5.1 General description

Our system can assume either of the two possible states: detection and tracking. In detection state, the system is busy searching all frames of the input video for a neutral face (front view of the face with closed straight lips and open eyes) and its relevant elements. In this state, no any information about detected face is known. Once a neutral face is found, the system switches to tracking, which means that it follows the changes in the position of the face and its elements and calculates the respective FAP values. In this state system uses information about face geometry and other face property estimated in detection state when a face was neutral. Using the information allows to achieve more stable tracking and higher FAPs accuracy. While in tracking the system analyses the face and the tracked changes at each moment, so that if tracking is not successful it must switch itself back to detection state (see Figure 4).

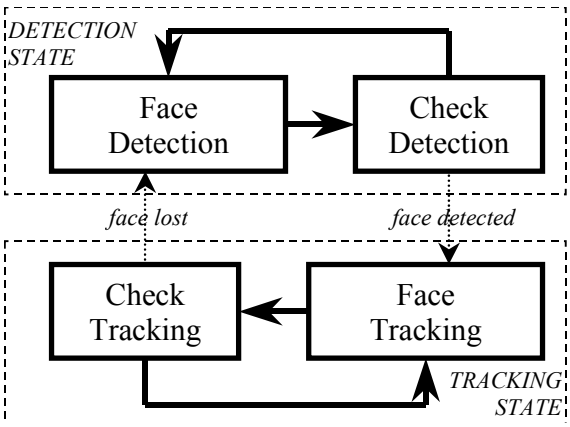

Figure 4. General scheme of video processing

5.2 Tracking state

On the tracking, state the system tries to track face position (two eyes points and two mouth points) and estimate FAPs of face on current frame. Such tracking procedure successfully tracks all face

movements and rotation for different lighting conditions but has several restrictions. In addition, this algorithm is not good to track faces with spectacles, moustaches and beard. The C realization of this algorithm showed that for processing 320x240 frame 2.2GHz Intel® Pentium (R) 4 based computer takes about 3-4 ms, it means that track speed is 200-300 FPS.

After eyes and mouth positions were tracked the FAPs are estimated. There are two different ways for FAPs estimation. The first makes rough FAPs estimation but does it in real time and can be used in on-line mode. The second one based on deformable templates techniques for eyes, lips and teeth tracking allows making more precise FAPs estimation but takes more computer time, so it can be used in off-line mode for pre-recorded video films processing.

The "off-line" estimation spends 260ms per frame for mouth processing and 130ms per frame for eyes processing on 2.2GHz Intel® Pentium (R) 4 based computer.

The "off-line" version is more exact and robust and consequently more perspective. In the same time it requires on the one hand more powerful processors and on the other hand algorithms and code optimization.

Both "on-line" and "off-line" pipelines are realized in FaceTA library. The "on-line" pipeline can process around 25-30fps and can be used to process video directly from video camera. The "off-line" pipeline can process around 2-2.5fps and so can be used only to process pre-recorded video.

6. FAPS CODING

In our pipeline MPEG-4 coding for a sequence of 16 frames applying one-dimensional discrete cosine transform to the whole set of FAP values is implemented using quantization and Huffman coefficient coding.

Two coding modes: intra and predictive are allowed. Note that all FAPs coding\decoding functionality is realized as a part of Intel Media Processing Library (MPL).

7. ANIMATION

In compliance with MPEG-4 standard face animation is specified by the frame-to-frame varying positions of 3D vertices on the scene object's surfaces and controlled by standard set of FAPs – facial animation parameters. Every low-level FAP is responsible for prescribing of the pointed facial feature point movement or the scene object transformation.

On-Line animation uses automatically calculated animation rules for given model in order to interpret FAPs and calculate new coordinates of model vertices. Finally we implement special post processing of mouth area since MPEG-4 standard claims to use for head animation only Facial Animation Parameters that makes strong FAP's coordination to be necessary especially for feature points located on natural head contour lines, such as inner and outer lips contours. We correct mouth area deformation during real-time animation to ensure mesh quality in mouth area and coordinate lip's motion.

For MPEG-4 visemes and expressions modeling physical-based muscle model was implemented. It uses line and circular muscles to model skin deformation in the mouth region. Muscle model provide calculation of smooth and life-like displacements for skin vertices. The input data for the muscle modeling procedure are the set of muscles modeled, the output are the calculated displacements for all skin vertices that are involved in motion for viseme or expression. As muscle modeling is slow enough, the use of muscle model is optional and offline.

Animation functionality is encapsulated in IFAL library.

Figure 5 allows estimating the quality of calibration and animation algorithms for different models.

Figure 5. Top line – personalized animated models; bottom line – their prototype person's photos

8. VISUALIZATION

Model rendering is the final stage in the pipeline. In the input each frame is a 3D face model deformed in accordance with FAPs while the output is a synthetic video showing a texturized "talking head", lit and projected on the screen. Visualization functionality is encapsulated in IFAL library and uses OpenGL for 3D model rasterization. Note that MPEG-4 doesn't specify any visualization approaches. The same model could look differently in various MPEG-4 players. To add to the realism of the model while rendering the system runs:

- 3D eyelash modeling and imaging,
- special eyes visualization for realistic pupil dilation
- physically based mouth illumination
- natural hair translucency modeling.
- expression wrinkles modeling. These, for example, if a smile is concerned, run from the nose to the corners of the mouth.

Figure 6. demonstrates our final result: the synchronized real and synthetic video playback.

Figure 6. Synthetic and real video.

9. PERFORMANCE ISSUES

The application developed to compute and encode a real video based FAP sequence (not precise online version) into a mpeg-stream showed the output productivity of 20-22fps, while the application to decode the FAP stream, animate and render the model shows productivity of 25 fps on P-IV 2.2GHz, 512 Mb RAM with NVIDIA Quadro4 board. It's enough for videoconferencing if common facial expression recognition only is necessary. But CPU speed increasing about 10 times is necessary to make precise facial features recognition & tracking on-line.

10. CONCLUSION

Full-automatic MPEG-4 compliant facial animation pipeline was developed.

Its current version provides a number of methods for automatic facial region detection, feature points recognition and tracking together with further FAP calculation based on real video sequence.

Two versions exist: more precise off-line and less precise on-line. The "off-line" version is more exact and robust and consequently more perspective. In the same time it demands more powerful processors and at the same time algorithms and code optimization.

FAP stream coding is implemented in full conformity with MPEG-4 standard specifics. Completely operational is the model-independent animation rules automatic computation method for the polygonal face model.

We developed a complete calibration pipeline that allow for the adjustment of a polygonal model of a generic head based on a set of photographs or video stream.

The all methods developed are ready for delivery and could be used team wise or independently in a lot of different applications – videoconferencing, web-based services, virtual characters driving, etc …

At the same time we faced the challenge of some MPEG-4 standard specifications restrictions, disadvantages and ambiguities that makes impossible to express real life facial gesture by FPs and FAPs only. Several techniques were successfully implemented to overcome these disadvantages.

Our results indicate suggested approach efficiency. However the task of full photo-realism achievement for Talking Heads still remains open research area cause such subtle thing as synthetic video quality estimation is not standardized and is strongly connected with peculiarity of human perception.

ACKNOWLEDGEMENTS

Special thanks to D. Ivanov and Computer Graphics Group of Moscow State University (Russia) and O. Mindlina and A. Pleskov (Intel Russia Research Center) for active help in work over the project.

11. REFERENCES

[1] SNHC, "*INFORMATION TECHNOLOGY – GENERIC CODING OF AUDIO-VISUAL OBJECTS Part 2: Visual*", ISO/IEC 14496-2, Final Draft of International Standard, Version of: 13, November 1998, ISO/IEC JTC1/SC29/WG11 N2502a, Atlantic City, October 1998.

[2] A.M. Tekalp, J. Ostermann "*Face and 2-D Mesh Animation in MPEG-4*", in Image Communication Journal, Tutorial Issue on the MPEG-4 standard http://leonardo.telecomitalialab.com/icjfiles/mpeg-4_si/8-SNHC_visual_paper/8-SNHC_visual_paper.htm

[3] http://www.biovirtual.com/

[4] D.Ivanov, V.Lempitsky, A.Shokurov, A.Khropov, Y.Kuzmin "*Creating Personalized Head Modesl from Image Series*", in Proc. of the International Conference on Computer Graphics & Vision GraphiCon'2003, Moscow, Russia.

[5] T. Firsova, D. Ivanov, V. Kuriakin, E. Martinova, K. Rodyushkin, V. Zhislina, "Life-like MPEG-4 3D "Talking Head" (beyond standard)", in Proc. of the 5th Int. Conf. On Computer Graphics and Artificial Intelligence 3IA'2002, Limoges (France), May 2002.

[6] The Virtual Reality Modeling Language - http://www.web3d.org/Specifications/VRML97/

[7] www.cacheforce.com.