REAL-TIME VISION AND SPEECH DRIVEN AVATARS FOR MULTIMEDIA APPLICATIONS

O. Schreer, R. Englert, P. Eisert, R. Tanger

***Abstract*—Recent progress in advanced video communication services and multimedia applications is grounded on novel human machine interfaces, improved usability and user friendliness driven by user centric research and development. In this paper, we describe a complete system concept and algorithmic details of an example application within this area. The key features of the system are vision and speech based interfaces, which are used to animate an avatar for an audio-visual representation of a communication partner. The system is applied in two application scenarios, namely video chat and customer care services. Both applications are mass-market oriented and therefore careful design and development of robust and supporting user interfaces are required. The presented approach is integrated into a complete real-time prototype system, which is permanently demonstrated in the showcase at the head quarter of Deutsche Telekom in Bonn, Germany.**

***Index Terms*—Avatar, multimodality, real-time tracking, segmentation**

I. INTRODUCTION

THE development of multimedia applications needs to take into account growing complexity and capability of interfaces, while the human ability to cope with them is limited. Thus the gap between systems' capabilities and human ability to make proper use of them is widening. A promising solution is to utilise emerging technologies in order to enable a more intuitive, more natural, user centric, human machine interaction and – communication. This can be done by providing the user with interfaces that fit to the specific user needs in their current usage context – and to enable the user to choose how to most appropriately interact with the system in a multimodal manner. Multimodal systems extend the so far established interaction options and thus enable new usage and service possibilities. Main driver to utilise a multimodal approach is the idea to empower a user to choose the most appropriate in- and output channel according to the specific usage situation and the user's task that is to be solved [1]. As prototype scenario an exemplary implementation of a multimodal avatar-based communication solution is chosen. The prototypical implementation investigates and demonstrates the possibilities, benefits and advantages of a multimodal user interface for certain usage situations and communication cases which are too complex, or have sophisticated requirements such as privacy protection.

The paper is organised as follows. In the next section, a prototype demonstrator of a multimodal avatar animation application is presented. We describe the two application scenarios, some state-of-the-art fundamentals on avatar animation in general and lip animation in specific. Some additional aspects on integration conclude this section. The focus of the paper is on real-time video analysis, hence, in section III, the skin-colour segmentation approach, the hand and head tracking module and the developed gesture recognition approach are described. These modules provide the desired animation parameters for real-time animation of the avatar. In section IV, we present details about the system performance and usability.

II. EXPERIMENTAL PROTOTYPE SYSTEM

A. System Overview

To demonstrate the above considerations of a multimodal user interface, we present an avatar animation system, which constitutes the core module of two new commercial advanced video communication applications. The first one is a call centre application, where the customer is able to communicate with an operator via the Internet. In contrast to conventional customer care services, the operator is also represented visually by an artificial character, the avatar. It is animated based on the live motion and the speech of the operator. Hence, the communication is enhanced by visual cues without transmitting live video streams. The latter one is often not desired due to privacy protection reasons. The second

Manuscript received January 30, 2007. The work is funded by Deutsche Telekom Laboratories, Germany.

O. Schreer is with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Einsteinufer 37, D-10587 Berlin, Germany (phone: ++49 30 31002 612; fax: ++49 30 3927200; e-mail: Oliver.Schreer@hhi.fraunhofer.de).

R. Englert is with the Deutsche Telekom Laboratories at Ben Gurion University, POB 653, Beer-Sheva 84105, Israel (Phone: +972 8 642 8115, Email: roman.englert@telekom.de).

P. Eisert is with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Einsteinufer 37, D-10587 Berlin, Germany (phone: ++49 30 31002 614; fax: ++49 30 3927200; e-mail: Peter.Eisert@hhi.fraunhofer.de).

R. Tanger is with the Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut, Einsteinufer 37, D-10587 Berlin, Germany (phone: ++49 30 31002 224; fax: ++49 30 3927200; e-mail: Ralf.Tanger@hhi.fraunhofer.de).

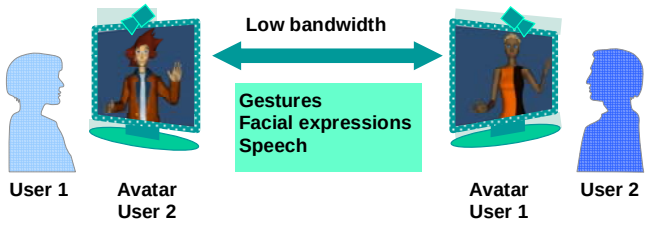

Fig. 1. Block diagram of the call centre application.

application is a chat scenario, where different people can meet in a virtual chat room. All chat partners are represented by an avatar, which is animated based on live motion and speech of each chat partner. In Fig. 1, the two way scenario is depicted.

For both applications a number of challenging requirements turn out from user perspective. At first, the hardware requirements must be as small as possible. Therefore, the application should run on conventional PC's without any dedicated hardware. A low cost Web camera and a standard head set must be sufficient to provide the required video and audio data. Additional challenges occur regarding usage of the application. The first is robustness under general working conditions, which includes changing lightning, arbitrary stationary or moving background and arbitrary clothing. Another challenge is related to the user, who has different knowledge about the system and its operation or who behaves unexpectedly. Specifically, vision-based systems and services developed for the consumer market may not claim any knowledge of the user neither on video technology nor on algorithms. Therefore a robust and user friendly application must consider many different issues in order to get accepted by the market. As shown in section III, many of these aspects are tackled by our prototype demonstrator.

The presented system fulfils completely the above mentioned hardware requirements. In Fig. 2, the block diagram of the call centre application is shown. The operator on the sender side is captured by a Web camera mounted on top of the display. Based on the video information, the position of the hands and the head orientation are tracked and converted to standard facial and body animation parameters as standardised in MPEG-4. Furthermore, a number of specific hand gestures shown by the user are recognised and used to enhance the animation capability of the avatar. In addition, the voice is captured and the audio signal is transmitted to the customer at the receiving side. The voice is analysed and visemes are generated to animate the lip shape corresponding

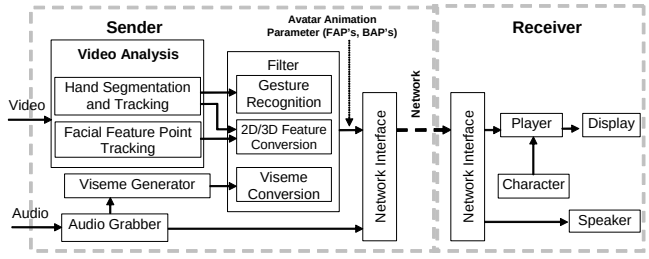

Fig. 2. Block diagram of the call centre application.

to different sounds. Visemes are the visual correspondence of phonemes and describe the shape of the lips while a certain phoneme is uttered. Beside the depicted modules in Fig. 2, additional features for avatar animation are implemented. For instance, while loading predefined sequences of motion parameters, the operator can activate specific high-level feature animations like opening sessions, leave-taking scenes or pointing gestures. As it is difficult to extract facial expressions from live video images the user can choose facial expressions manually. The following set of facial expressions is available: neutral, joy, sadness, anger, fear, disgust and surprise. These additional animation features can be applied by using the GUI of the application, where specific buttons are assigned to activate animation sequences or facial expressions.

All head and hand motions, facial expressions and lip movement are described via body animation parameters (BAP) and facial animation parameters (FAP) according to the definition in the MPEG-4 standard. The complete set of animation parameters and the audio signal are then transmitted to the customer on the receiving side. As no video information is necessary, this approach is efficient in terms of the required bandwidth and therefore appropriate in web-based customer care applications.

B. Avatar Animation

For this communication scenario, avatars are exploited for the representation of the participants. In order to animate a synthetic character, the static shape must be defined as well as the dynamic motion. The appearance of an avatar can be modelled by image-based rendering techniques [2], but here a polygonal surface mesh is used which is coloured by mapping texture maps onto the surface. Fig. 3 shows an example of an avatar by means of a triangle mesh (left) and the corresponding coloured version after applying texture maps (right). Rendering a textured mesh can be performed in real-time with very little effort due to the hardware acceleration on the graphics board, but nowadays even devices without hardware graphics support can render avatars in real time.

More interesting than the static modelling of the avatar is the description of global and local motion/deformation in order to represent body motion or facial expressions. To achieve full control over the behaviour of the synthetic character, a parameter-based method is used. Each parameter affects a small elementary and locally restricted motion and the actual appearance of the character is defined by superposing all basic motions. The shape can thus be defined for each time instance by a vector of animation parameters.

Fig. 3. Wireframe representation of an avatar and coloured version.

For faces, an early approach is the Facial Action Coding

System (FACS) [3], which defines 46 different groups of muscles in the face that can be moved independently of each other. Rather than having a medically motivated modelling of facial expressions, ISO/MPEG standardised a description of facial motion based on a pure visual appearance [4][5]. Instead of starting from groups of muscles that can affect also larger regions in the face, the facial animation parameters (FAPs) control only very local surface deformations, as, e.g., movements of the lip corners, eye, eyebrows or cheeks. A total set of 66 different parameters define the current facial expression. Fig. 4 shows some examples of movements that can be created by varying the set of basic facial animation parameters. We have also implemented some high level expressions and visemes. The viseme information is directly determined from speech as described in the next section so that lip movements can be created synchronously to the audio signal. In contrast, other facial animation parameters like head rotation or eye blinking are derived from the live video input.

Similar to the facial animation parameters, in MPEG-4 also body motion parameters BAPs [4] are defined to control the movements of arms, legs, and other body parts. While the FAPs define surface deformation in the face, the 186 BAPs, based on the H-anim standard, mostly define rotation of all the joints of the skeleton. The effect of bone movement on the outer tissue is not defined and has to be handled by the individual player implementation. For our system we also added a subset of the MPEG-4 body animation parameters, to enable gestures of arms and fingers as illustrated in Fig. 4.

Although not available in MPEG-4 we have also added some high level features for body motion similar to the expressions in the face. Especially for the hand with the large number of different joints some high level gestures (pointing, ok, victory, etc.) are specified which are controlled by a single parameter. These more sophisticated motion structures are internally mapped on the larger number of BAPs.

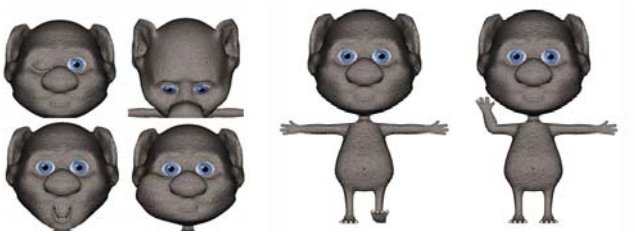

Fig. 4. Face and body motion controlled by MPEG-4 FAPs / BAPs.

C. Speech Analysis for Lip Animation

Avatars cannot be only animated from video input. The high correlation between lip movements and the spoken words allows the derivation of facial animation parameters also directly from the speech signal. In some cases, the video input can even be totally omitted reducing computational complexity and bit-rate for transmissions of animations over networks. Researchers have therefore been exploiting audio-visual correlation for a long time. Early approaches [6][7] used lip motion derived from audio signals to enhance coding efficiency of head-and-shoulder video sequences. With the increase in 3D rendering quality, also avatars have directly been controlled from speech [8][9][10] in applications like video rewrite, human-machine-interfaces, or electronics greeting cards.

In order to estimate lip motion from speech, the signal is usually processed in blocks, assuming constant properties within each block. In this approach, windows of 20ms length are processed every 10ms. From the large number of audio samples, a much smaller number of features are computed, which are then used to classify the phoneme corresponding to the audio block. It is desired to have a small number of features which still contain the information about lip shape. On the other hand, noise and speaker dependent information like the pitch should be removed to increase robustness and generality. Different features are used in literature like linear predictive coding (LPC), formants, cepstral coefficients, or MFCCs (Mel Frequency Cepstral Coefficients). In addition, features like zero crossing rates can provide more information, for example about fricatives. All these features are combined into a single feature vector having about 10 to 20 elements. From each vector, a corresponding phoneme is estimated using different types of classifiers. Neural networks [11] can map between feature vectors and lip shape. Hidden Markov Models (HMMs) are used to incorporate also temporal statistics between the small audio segments, increasing robustness of the phoneme recognition. In this approach, we also exploit an HMM-based recogniser.

In speech recognition, the goal is to estimate a sequence of phonemes with very high accuracy, using a-priori knowledge about a vocabulary or a language. For animating an avatar, less precision is required simplifying the entire recognition task. Here, visemes corresponding to a particular mouth shape have to be estimated instead of phonemes with the number of visemes being much smaller than the number of phonemes. 15 different visemes are used to display all analysed phonemes. In the case of displaying a specific facial expression such as anger or happiness, the animation parameters of the visemes overwrite the actual FAPs relevant for the mouth shape.

D. Integration aspects

To allow a flexible and easy integration of several prototypes with different configurations the system integration follows a data driven module concept. Modules have generic inputs and outputs called “pins” which accept video as well as audio or animation parameters depending on the data a particular module needs or provides. Because of the high demand on computation power all modules have the ability to run asynchronously in their own thread. This allows the full exploitation of modern multiple core CPUs.

Fig. 5 shows the core server components of both prototypes. Results of video and audio analysis are first described in a syntax independent from the currently used avatar player. The animation parameter generation module is responsible for the transformation of all analysis results as well as the submitted high level expressions into the particular

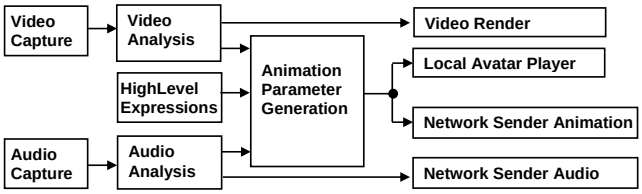

Fig. 5. Core server modules.

format needed by the current player. The video render module allows direct observation of video analysis results and provides a visual feedback for parameter adjustment. After applying the animation parameters the fully animated "own" avatar can be observed in the local avatar player. Thus the user always has full insight of what happens with his avatar. Animation information and audio are transmitted over the network using standard UDP or TCP connections.

III. Video Analysis of Human Gestures

The animation of avatars based on the live motion of the user is a great challenge even in the case of a system for everybody's use. Specifically, services developed for the consumer market may not claim any knowledge of the user neither on video camera technology nor on algorithms. The user may not perform any initial gestures or specific positions in order to start the processing and it must be allowed to behave natural, while using the system. Furthermore, the surrounding environment should not underlay specific constraints in terms of the background, moving objects or lightning conditions and the algorithm must be user independent. These are challenging constraints, which are not completely fulfilled by current vision-based tracking and animation systems. Tracking of human bodies and faces as well as gesture recognition has been studied for a long time and many approaches can be found in the literature. A survey on human body tracking is given in [12]. Hand gesture recognition is reviewed in [13] and a 3D gesture recognition system is presented in [14]. Tracking the user's face and estimating its pose from monocular camera views is another important issue. As the 3D information is lost during perspective projection onto the image plane, some model assumptions have to be applied in order to estimate the 3D pose. In [15], some specific face features are tracked in order to recover the orientation and position of the user's head. In the considered scenario of animating a virtual human, the accuracy of 3D positions of head and hands does not play that important role, but the immediate transfer of general live motion to the virtual human is required such as waving hands, pointing gestures or nicking the head. This allows some simplifications in terms of accuracy, but introduces additional challenges regarding smoothness and reliability of the animated motion. In this section, we will present a real-time system, which covers a lot of the above mentioned challenges and dispense with specific knowledge of or behaviour by the user. The system is capable of transferring the motion of hands, representing nick and shake of the head and also recognising gestures. The whole video analysis module can be subdivided in the following parts (see Fig. 6):

- estimation of skin-colour segmentation parameters
- initial blob detection
- tracking and segmentation
- facial feature tracking for head rotation animation
- gesture recognition

Numerous applications use skin-colour as one of the basic features for detecting and tracking of human face and hands. They have different aims and different constraints under which the human face or hands are being analysed. Example applications dealing with segmentation of hands are hand sign recognition, human vehicle interaction or human computer interfaces [16][17][18][19]. In some hand segmentation approaches marked gloves are used, which are not applicable in video communication systems [20]. In other approaches infrared cameras are used or depth information based on multiple views is exploited [21][22][23]. The real-time constraint is considered as well in gesture recognition applications for tele-manipulation, virtual reality and other human-computer interaction applications [24][25][26][27].

In the next subsections, we will discuss the different modules in more details and reference to previously mentioned challenges of a robust and user friendly real-time

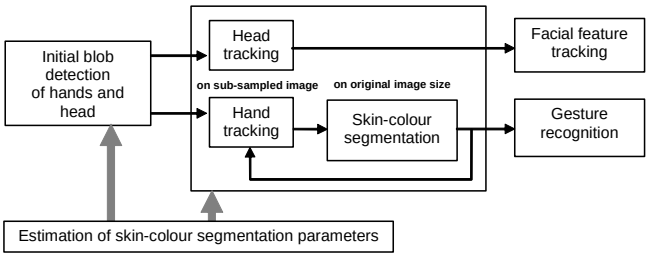

Fig. 6. Block diagram of segmentation and tracking.

application.

A. Skin-Colour based segmentation and estimation of segmentation parameters

In many former applications skin colour turned out to be a very helpful feature for detecting and segmenting hands and faces. Hence, the human skin-colour can be defined as a "global skin-colour cloud" in the colour space [28]. The general thresholds for this striking area are still too large to obtain reasonable segmentation results. Depending on different factors like shadows, illumination, colour distribution in the particular video data, different pigmentation of the person's skin and so on, it is useful to adapt thresholds to the given illumination conditions and the observed person. Hence, an important question arises how to determine appropriate skin-colour parameters for a scenario to achieve best segmentation results. Applying parameters from general statistical analysis of skin-colour does not lead to optimal segmentation results in the majority of cases. But they can be often used as coarse start values to find appropriate parameters by slightly varying them.

One option is to adapt the thresholds manually at start of

the application. Obviously, this is not convenient in terms of usability and user friendliness. Therefore, a quasi-automatic method is desired to find suitable parameters. Usually, TV- and video data are available in the YUV-colour space. Hence, the investigations have been made in the YUV-colour space. In our approach we consider the chrominance (U, V channel) representing the colour information. In addition to that, the range of luminance values must be restricted even in the case of very bright or dark regions. The skin-colour is then described by the mean values m_y, m_u and m_v and the tolerance values σ_y, σ_u, and σ_v. The tolerances of the U, V-channel are relatively narrow, compared to a very large luminance range.

The quasi-automatic procedure for estimation of the actual segmentation parameters performs at program start, where the user is asked to move its hands. Due to this, the search range for possible skin coloured regions can be limited to the moving area. Based on predefined tolerances, the mean values of Y, U and V are changed according to an optimisation criterion, which is defined as the most significant large and closed region. Since the tolerances are kept fixed and the whole optimisation is applied on the subsampled image, the parameter variation performs reasonably fast and provides the optimal set of segmentation parameters.

B. Initial blob detection

The main challenge for initial blob detection is the immediate localisation of hands and head in arbitrary conditions and without any constraints on user behaviour. This includes also people wearing t-shirts, where skin coloured regions are not only limited to hands. Based on the estimated segmentation parameters from the previous section, a segmentation of the subsampled image is performed to reduce the processing time. The general approach is to analyse the row and column histogram of the binary segmentation result, whereby reasonable blobs are searched for. In this module, the assumption is made, that the head blob occurs in the top of the image. After finding the head blob, the left and right hand are detected, if visible. In order to distinguish between different skin-coloured regions in the case of wearing t-shirts, motion information is exploited. The combined result of motion detection and skin-colour segmentation provides unambiguously the desired hand blobs. In Fig. 7 results of motion detection, skin-colour segmentation and combination of both are shown. These results are obtained from the most right original image showing the resulting segmented hands.

The initialisation is performed and segmentation and tracking are started as soon as the head and at least one separated skin-colour blob are detected. If the hands get lost during tracking or segmentation, the head position is exploited to re-initialise immediately the system by evaluating the row and column histogram omitting the head area.

Fig. 7. Result of motion detection, skin-colour segmentation, combination of both and final tracking result of detected hands (from left to right).

C. Tracking and segmentation of head and hands

The main principle of this module is a closed tracking and segmentation loop. Each centre position of already segmented hand, the so-called seed point, in the previous image is checked, if it still lies in a skin-coloured region of the current image. If this is the case, the seed point is used for accurate segmentation on full resolution, which is achieved by applying a region growing technique. Starting from the seed point of the hand, the segmented area is enlarged by analysing continuously the neighbours of the segmented pixels. The result is a closed region of high accuracy. The centre of gravity of this segmented hand region is used as seed point for the next video frame. The precise segmentation result of the bounded region is used for the subsequent gesture recognition module later on. Should the initial check of the seed points fail (e.g. caused by holes in the segmentation result), the algorithm searches for skin-coloured areas in the surrounding region. If this fails as well, then, the initial blob detection is applied in order to search for significant skin-colour blobs in the whole image. The tracking and segmentation module also considers overlap between both hands and the head. More details can be found in [29]. In Fig. 8, a successful tracking in the case of overlapping regions is shown. The approach achieves real-time performance due to tracking on subsampled images and skin-colour segmentation limited to bounding boxes circumscribing the hands and the head.

D. Facial feature tracking for head rotation

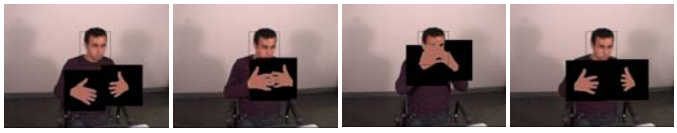

Fig. 8. Sequence of critical hand/head overlap.

The aim of our facial feature tracking is to derive a convincing and reliable rotation of the user's head. In contrast to semantic facial feature tracking approaches [30], the proposed solution is based on analysis of the relative motion of a few but robust general feature points in the face. A common feature tracker is applied on the skin-coloured pixels inside the bounding box of the user's head. The feature tracker is based on a two-step approach. First, relevant features are selected by using a corner operator e.g. Harris detector. Secondly, the selected features are then tracked continuously across frames by using a dissimilarity measure. This guarantees, that features are discarded from further tracking in the case of occlusions. Even in the case of a rotating head some good features become invisible and get lost.

In Fig. 9, the features are shown in the face region for three successive frames. The big cross assigns the centre of all skin pixels. The considered skin colour region is marked by the line around the face. Due to the blond hairs of the test person, the hairs are recognised as well as skin. Based on a few

Fig. 9. Facial feature tracking result of three successive frames.

robustly tracked facial features, the head orientation can be derived by comparing the relative motion of facial features to the projected 2D motion of the head. In the case of a horizontal or vertical rotation the motion of the mean of all face pixel positions is significantly smaller than the relative motion of the facial features. This behaviour of facial feature points allows a simple approximation of the head rotation in horizontal (turn angle) and vertical direction (nick angle). The median value of horizontal and vertical coordinates of facial feature points is assigned to $(\overline{m}_t, \overline{n}_t)$, whereas the mean of all face pixel positions is denoted by $(\overline{p}_t, \overline{q}_t)$. The relative change of facial feature points (horizontal/vertical) is then calculated by (1) and the change of horizontal and vertical rotations is approximated by (2). A scale factor γ is introduced to adopt the pixel unit to angles.

$$\Delta u = (\overline{m}_t - \overline{m}_{t-1}) - (\overline{p}_t - \overline{p}_{t-1}), \Delta v = (\overline{n}_t - \overline{n}_{t-1}) - (\overline{q}_t - \overline{q}_{t-1}) \quad (1)$$

$$\Delta\varphi_u = \sin(\gamma \cdot \Delta u), \quad \Delta\varphi_v = \sin(\gamma \cdot \Delta v). \quad (2)$$

As it is obviously not possible to calculate the absolute rotation from this method, drift effects may occur. This can be avoided by continuously weighting the current turn (or nick) angle by some factor smaller than 1. As the central viewing direction is the most relevant one, the animated head will adopt to this position after a while.

E. Gesture recognition

The skin-colour based segmentation algorithm provides a very accurate silhouette of the user's hand, which allows robust recognition of many typical gestures. If one of these gestures is shown by the user and recognised by the system, it can be immediately transferred to the avatar on the receiving side. In contrast to template based approaches e.g. [31], we applied a method based on evaluation of a specific distance function derived from the contour of the hand segment. This approach has been applied successfully in other systems [32], but in contrast to that proposal the gesture can be shown quite naturally without posing it on a specific flat panel. The main approach is to detect the number of fingers, their position and their orientation, which provide sufficient information in order to recognise many basic gestures from the American Sign Language (ASL) alphabet [33]. The chosen gestures are the most distinguishable ones related to their surrounding contour. For each contour point, the normal of its tangent along the contour is calculated. Then, following this normal, the distance to the contour point at the opposite side of the contour is measured. In the resulting distance function, the peaks assign clearly the fingers in the silhouette but also valleys between neighboured fingers. In order to distinguish between both types of peaks, the curvature is used as selection criteria. The resulting peaks are assigned to fingertips and based on the orientation of the fingers a classification into many different gestures can be performed. In the current implementation 13 different gestures are recognised robustly. In Fig. 10 and Fig. 11, the silhouette of a two and three finger gesture is shown on the left. The black lines assign the resulting orientation of the detected fingers. On the right hand side of both silhouettes, the related distance function is depicted, which illustrates the capability for successful classification of hand gestures.

Concerning a convincing animation, the key challenge in this task is twofold: Gestures, shown by the user must be identified correctly in a very short time. On the other hand, arbitrary hand positions should not be interpreted as gestures. Hence, the aim is to solve a standard classification problem, which is the reduction of false negative and increase of true positive events. This must hold for the whole session. In order to reduce the false negatives, a gesture is transferred to the avatar, if the following two temporal conditions are fulfilled:

1. In general, a gesture is performed without moving the hand. Hence, the motion of the hand is assumed to be small in the case of a shown gesture.
2. A gesture must be recognised among a number of succeeding frames accepting a few false detections.

If these conditions are fulfilled, a gesture shown by the user can be assumed, as depicted in the example screenshots in Fig. 12. More details can be found in [34].

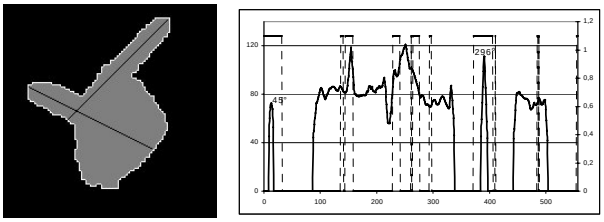

Fig. 10. Silhouette of a pointing gesture (left) and related distance function (right).

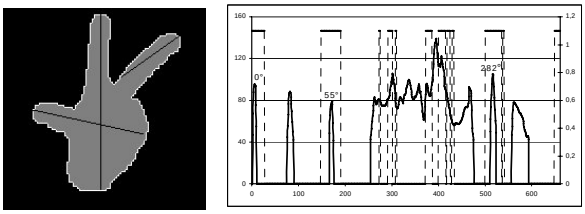

Fig. 11. Silhouette of a three-finger-gesture (left) and the related distance function (right).

Fig. 12. Example screen shots showing two different gestures.

F. *Computation of animation parameters*

The results of hand and head tracking as well as gesture recognition have to be converted to specific animation parameters of the avatar. The positions of the hands are derived from the result of the skin-colour segmentation and tracking module described in section III.B. It provides reasonable and stable results of the motion of both hands in the 2D image plane. To achieve natural avatar movements, the 2D positions have to be transferred onto the 3D model of the avatar. Since only two degrees of freedom of the hand position are available, just a simplified motion model is implemented based on the assumption that the hands of the avatar mainly move within a 2D plane in the 3D space. Thus, taking into account some further physical constraints such as the restricted range of elbow joints and the proportions between the upper arm and the forearm, the 3D position of the avatar's hands can be computed from these 2D tracking results. The resulting avatar motion is obviously not exactly the same as the 3D motion of the user's hand. This is fully acceptable, because the main aim is to provide a realistic, natural avatar motion.

The head rotation such as head nick and turn can be derived easily from the module described in section III.D. Again, the aim is not to reconstruct the correct head orientation, but to represent meaningful head gestures like head nick and turn. The result of gesture recognition is transferred to the avatar by moving the arm in a predefined position and showing the recognised gesture by changing the fingers of the specific hand accordingly. The hand tracking and gesture recognition is performed independently for the left and right hand. Due to this type of application, it is not required to reconstruct exactly 3D positions of hands or absolute head orientations. The main challenge is to provide the user with a smoothly animated avatar which behaves as natural as a human being.

IV. System Performance

The presented prototype system runs on a state-of-the-art Laptop at 25 frames/s. The video capture unit is a standard off-the-shelf wide angle web camera, which is mounted on the top of the display. The wide viewing angle camera allows the user to sit at a normal position in front of the laptop, whereby the full upper body is captured. Due to this, the user is allowed to move the arms in the whole viewing space without any restriction. Because of the skin-colour-based segmentation and tracking scheme, skin-coloured background or clothes need to be avoided. Nevertheless, the tracking scheme is robust regarding users wearing t-shirts, as described in section III.B. The initialisation of the tracking and segmentation module performs automatically as soon as a skin-coloured blob is detected beside the head. There is no specific user behaviour required and therefore the response of the system i.e. the animation result is perceived immediately.

The presented gesture recognition approach is based on the highly accurate skin-colour hand tracking and segmentation algorithm. The resulting silhouette is exploited to derive many different features, which allow the classification into a large set of finger gestures. Currently, 13 different gestures from the American Sign Language are implemented. The approach does not require any training set or database of gestures, because it is completely data driven. Hence, the proposed algorithm is independent on the user. Due to the defined temporal conditions during gesture recognition, a delay of about 300ms is introduced, but it is not noticeable by the user. Therefore, the current recognition performance is also acceptable for interactive human-machine interfaces.

During algorithm development, 12 test persons evaluated the gesture recognition module, its recognition capability and the usability of the system. They were not familiar with the application and therefore completely inexperienced. In individual sessions, the test persons had to test the feedback by the system concerning the gestures they presented. After the session, they have been asked how easy or difficult it was to receive the correct feedback from the system regarding correct recognition of the gesture. An evaluation scheme similar to the mean opinion score has been used for this evaluation. A score of 5 represents very good recognition by the system, whereas a score of 1 assigns a very poor recognition capability. The users have been asked to present each of the gestures from a set of 13 different gestures as shown in Fig. 13 (left). In the diagram in Fig. 13 (right), the evaluation result is presented for all 13 gestures. Just a minor number of gestures receive a lower score. The reason for this depends on a weak differentiation between some gestures e.g. one finger gestures, or on difficulties of the users to present

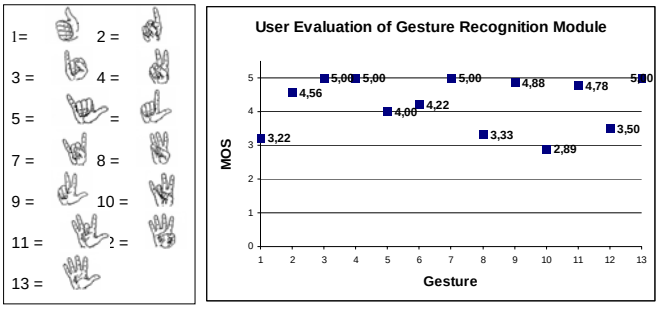

Fig. 13. A set of 13 gestures (left) based on the American Sign Language and MOS evaluation result of the gesture recognition capability (right).

them clearly such as gesture no. 10.

The system has been presented successfully in the demonstration area at European Conference on Computer Vision (ECCV) 2006 in Graz, Austria. The visitors were able to explore the application and the feedback received was positive. The prototype demonstrator is currently installed in the showcase of the head quarter of Deutsche Telekom in Bonn, Germany and at the Deutsche Telekom Laboratories head quarter in Berlin, Germany. Many presentations per day as well as the positive feedback by the visitors prove the novelty and the acceptance of such an application. The continuous user feedback is required to further improve the

overall system, to add novel features and to enhance the audio-visual user interface.

V. Conclusion and Future work

Multi-modal interfaces in human computer interaction become prominent in many existing application areas and new fields of application arise due to the new interface capability. We have presented a complete system concept and its realisation in a prototype demonstrator in the emerging field of video communication. Audio-visual information is used and analysed to animate an artificial avatar, which represents the communication partner in a reasonable natural way. The presented approach considers a complete system, from system and software design point of view to usability, user friendliness and robustness. The latter one is important since a commercialisation of such application is planned. The provided technology on human motion and gesture analysis can be used in many other application scenarios such as novel human-machine interfaces. Simple user actions or gestures can be interpreted by the system and may activate user commands. Integration in mobile devices is foreseen; where new modalities can be provided in addition to touch screens, stylus or speech recognition.

VI. Acknowledgement

We gratefully thank France Telecom R&D, for the provision of the avatar.

References

[1] R. Englert, G. Glass, “An Architecture for Multimodal Mobile Applications”, 20th Symp. on Human Factors in Telecommunication (HFT 2006), Sophia Antipolis, France, ETSI, 2006.

[2] P. Eisert, J. Rurainsky, “Geometry-Assisted Image-based Rendering for Facial Analysis and Synthesis”, Signal Processing: Image Communication”, vol. 21, no.6, pp. 493-505, July 2006.

[3] P. Ekman and W. V. Friesen. Facial Action Coding System. Consulting Psychologists Press, Inc., Palo Alto, 1978.

[4] ISO/IEC FDIS 14496-2, Generic Coding of audio-visual objects: (MPEG-4 video), International Standard, 1999.

[5] I. S. Pandzic, R. Forchheimer, “MPEG-4 Facial Animation - The standard, implementations and applications”, Igor S. Pandzic, Robert Forchheimer (editors), John Wiley & Sons, 2002, ISBN 0-470-84465-5.

[6] S. Morishima, K. Aizawa, and H. Harashima, “An intelligent facial image coding driven by speech and phoneme”, ICASSP, pp. 1795–1798, 1989.

[7] R.R. Rao, T. Chen, “Exploiting audiovisual correlation in coding of talking head sequences”, Picture Coding Symposium, pp. 653–658, Mar. 1996.

[8] B.J. Theobald, G.C. Cawley, I. A. Matthews, J. A. Bangham, “Near videorealistic synthetic visual speech using non-rigid appearance models”, Proc. International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 5, Hongkong, pp. 800–803, 2003.

[9] I.A. Ypsilos, A. Hilton, A. Turkmani, P. Jackson, “Speech-driven face synthesis from 3D video”, Proc. IEEE Symposium on 3D Data Processing, Visualisation and Transmission, 2004.

[10] Y. Chang, T. Ezzat, “Transferable videorealistic speech animation”, Proc. ACM Eurographics, Los Angeles, USA, pp. 29–31, 2005.

[11] P. Eisert, S. Chaudhuri, B. Girod, “Speech Driven Synthesis of Talking Head Sequences”, Proc. Workshop 3D Image Analysis and Synthesis, Erlangen, Germany, pp. 51-56, Nov. 1997.

[12] T.B. Moeslund, E. Granum “A survey of computer vision-based human motion capture”, Computer Vision and Image Understanding, vol. 81, no. 3, 231-268, 2001.

[13] I. Pavlovic, R. Sharma, T.S. Huang “Visual interpretation of hand gestures for human–computer interaction: a review”, IEEE Trans. on PAMI, 19, pp.677-695, 1997.

[14] A. Just, S. Marcel, O. Bernier, “HMM and IOHMM for the recognition of Mono- and Bi-manual 3D Hand Gestures”, British Machine Vision Conf., Kingston Univ. London, 2004.

[15] P. Eisert, B. Girod, “Analyzing Facial Expressions for Virtual Conferencing”, IEEE Computer Graphics & Applications: Special Issue: Computer Animation for Virtual Humans, vol. 18, no. 5, pp. 70-78, 1998.

[16] K. Imagawa, S. Lu, S. Igi, „Colour-Based Hands Tracking System for Sign Language Recognition“, Int. Conf. on Automatic Face and Gesture Recognition, pp.462-467, 1998.

[17] X. Zhu, J. Yang, A. Waibel, “Segmenting hands of arbitrary colour”, Int. Conf. Autom. Face Gesture Recognition, pp.446-453, 2000.

[18] L. Sigal, S. Sclaroff, V. Athitsos, “Skin Colour-Based Video Segmentation under Time-Varying Illumination”, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 26, No. 7, pp.862-877, 2004.

[19] T. Starner, B. Leibe, D. Minnen, T. Westyn, A. Hurst, J. Weeks, “Computer Vision-Based Gesture Tracking, Object Tracking, and 3D Reconstruction for Augmented Desks”, Machine Vision and Applications, Vol. 14(1), pp. 59-71, 2003.

[20] K. Dorfmueller-Ulhaas, D. Schmalstieg, „Finger tracking for interaction in augmented environments“, ACM/IEEE Int. Symp. on Augmented Reality (ISAR 2001), pp.30-44, 2001.

[21] K. Oka, Y. Sato, H. Koike, “Real-time tracking of multiple fingertips and gesture recognition for augmented desk interface systems”, Fifth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 411-416, May 2002.

[22] S. Malassiotis, F. Tsalakanidou, N. Mavridis, V. Giagourta, N. Grammalidis, M.G. Strintzis, “A face and gesture recognition system based on an active stereo sensor”, Int. Conf. on Image Processing, vol.3, pp.955-958, 2001.

[23] C. Jennings, “Robust finger tracking with multiple cameras”, Proc. of IEEE Int. Workshop on Recognition, Analysis & Tracking of Faces & Gestures in Real-Time Systems, pp.152-160, 1999.

[24] B.C. Lovell, D. Heckenberg, “Low-Cost Real-Time Gesture Recognition,” Proc. of Asian Conference on Computer Vision, pp.336-341, 2002.

[25] M. Hasanuzzaman, V. Ampornaramveth, Z. Tao, M.A. Bhuiyan, Y. Shirai, H. Ueno, „Real-time Vision-based Gesture Recognition for Human Robot Interaction“, IEEE Int. Conf. on Robotics and Biometrics, pp.413-418, 2004.

[26] R. Herpers, W. J. MacLean, C. Pantofaru, L. Wood, K. Derpanis, D. Topalovic, J. Tsotsos, “Fast Hand Gesture Recognition for Real-Time Teleconferencing Applications”, Int. Workshop on Recognition, Analysis and Tracking of Faces and Gestures in Real-Time Systems, pp.133-144, 2001.

[27] C. Costanzo, G. Iannizzotto, F. La Rosa, “VirtualBoard: real-time visual gesture recognition for natural human-computer interaction”, Proc. of Int. Parallel and Distributed Processing Symposium, p.112, April 2003.

[28] M. Störring, H.J. Andersen, E. Granum, “Skin colour detection under changing lighting conditions” Symp. on Intelligent Robotics Systems, pp. 187-195, 1999.

[29] S. Askar, Y. Kondratyuk, K. Elazouzi, P. Kauff, O. Schreer, “Vision-based Skin-Colour Segmentation of Moving Hands for Real-Time Applications” Proc. of 1st European Conf. on Visual Media Production (CVMP), London, United Kingdom, 2004.

[30] Y. Hu, L. Chen, Y. Zhou, H. Zhang, “Estimating face pose by facial asymmetry and geometry” Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recognition, pp.651-656, 2004.

[31] T. Coogan, G. Awad, J. Han, A. Sutherland, “Real Time Hand Gesture Recognition Including Hand Segmentation and Tracking”, Proc. of Int. Symp. on Computer Vision, ISVC06, 2006.

[32] G. Iannizzotto, F. Rosa, C. Costanzo, P. Lanzafame, “A Multimodal Perceptual User Interface for Collaborative Environments”, Int. Conf on Image Analysis and Processing, pp.115-122, Cagliary, Italy, 2005.

[33] Z. Mo, U. Neumann, “Lexical Gesture Interface”, IEEE Int. Conf. on Computer Vision Systems (ICVS '06)., pp.7, 2006.

[34] O. Schreer, S. Ngongang, “Real-time Gesture Recognition in Advanced Videocommunication Services”, Int. Conf on Image Analysis and Processing, Modena, Italy, Sept. 2007.