Real-time avatar animation steered by live body motion

Oliver Schreer[1], Ralf Tanger[1], Peter Eisert[1], Peter Kauff[1], Bernhard Kaspar[2], Roman Englert[3]

[1]Fraunhofer Institute for Telecommunications/Heinrich-Hertz-Institut, Einsteinufer 37, 10587 Berlin, Germany
{Oliver.Schreer, Ralf.Tanger, Peter.Eisert, Peter.Kauff}@fraunhofer.hhi.de
http://ip.hhi.de
[2]T-Systems International GmbH, Am Kavalleriesand 3, 64295 Darmstadt, Germany
Bernhard.Kaspar@t-systems.com
[3]Deutsche Telekom AG, Laboratories, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
roman.englert@telekom.de

Abstract. The future customer service provided by call centres will be changed due to new web-based interactive multimedia technologies. Technical support will be offered in a completely new way by using advanced image processing technologies and natural representation of virtual humans. We present a prototype system of an animated avatar, which is steered by live body motion of the operator in a call centre. The hand and head motion is transferred directly to the avatar at the customer side in order to support a more natural representation of the virtual human. The system tracks the operators hands and the head motion quite robust in real-time without specific initialization based on a monocular camera.

1 Introduction

Tracking human bodies and faces has received a lot of attention in computer vision research in the last years. The reason is, that a number of interesting applications have been raised in the past such as motion capture for entertainment industry or medical purposes, human-machine interaction, automatic surveillance systems or interactive web-based commercial applications. A lot of robust approaches have been developed, which are now going to be carried over to commercially available systems. Therefore, a lot of new challenges like robustness under different lightning conditions, independency from different users, eased use without sophisticated initialization procedures turn out. In this paper, a call centre application will be presented, where an operator is represented to the customer via an animated avatar. The head and body motion of the operator is immediately transferred to the virtual human by using robust skin colour segmentation and facial feature tracking algorithms. The complete image processing is performed on monocular colour video images in real-time on a conventional PC at full video frame rate.

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 [1]. A real-time body tracking system using

structured light without use of additional markers is presented in [2]. This constraint is particularly important in user-friendly applications. Hand gesture recognition is reviewed in [3] and a 3D hand gesture recognition system is presented in [4]. 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 [5].

In [6], some specific face features are tracked in order to recover the orientation and position of the users head. Other methods use IR illumination, which simplifies tracking of the eyes [7]. 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 in terms of smoothness and reliability of the animated motion. In the next section, the system of a call centre application is presented. Although this system also includes speech analysis, the focus of this paper is on image processing. Hence, the skin-colour segmentation and facial feature extraction is reported briefly. Based on the specific aims of this application, the reconstruction of the hand and head position and the head orientation is explained. Finally, results are shown and the article ends with a conclusion.

2   System overview

As shown in **Fig. 1**, the considered application provides for an operator on the sender side, who is captured by a video camera mounted on top of the display. Based on the video information, the position of the hands and the head orientation are registered and converted to standard facial and body animation parameters as standardised MPEG-4 (Part 2 (Visual)).
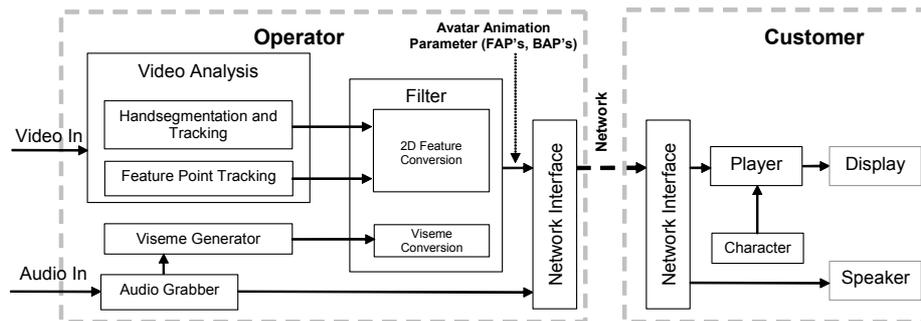
Fig. 1. Block diagram of the call centre application using an animated avatar

In addition, the voice is captured and the audio signal is transmitted to the customer at the receiving side. The captured voice is analyzed and visemes (a visual representation of phonemes) are generated to animate the lip shape corresponding to different sounds. Based on these visemes, a natural lip movement of the avatar can be

reproduced. Beside the depicted modules in **Fig. 1**, additional modules are implemented in order to provide more complex and natural animation. For instance, while loading predefined sequences of motion parameters from a library, the operator can activate specific high-level feature animations like opening sessions, leave-taking scenes or pointing gestures. If tracking and high-level feature animation are both switched off due to certain reasons, the natural behaviour of the animated avatar is improved by slight random motion of the head and parts of the face (e.g. eye blinking). General facial expressions like friendliness, anger or fear, which are difficult to extract from live video images can be chosen by an additional expression generator. 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. The customer is viewing a virtual human represented by an avatar, which is steered by the live body motion and speech. In the next two sections, more details will be presented regarding the image processing part of the system. The skin-colour segmentation algorithm that is used for tracking hand and head regions is explained first. Then, the algorithm, which derives the 3D position of the hands from the corresponding segments and which is used for steering the hand movements of the avatar, is presented. Subsequently, the algorithm for tracking head features is described, and it is explained how it is used for animating head rotation. Finally, some experimental results are shown and a conclusion ends the article.

3   Skin-colour segmentation and tracking

The colour of human skin is a striking feature to track and to robustly segment the operators hands and face. It is exploited, that human skin colour is independent on the human race and on the wavelength of the exposed light [8]. The same observation can be made considering the transformed colour in common video formats. Hence, the human skin-colour can be defined as a "global skin-colour cloud" in the colour space [9]. This is utilised successfully in a fast and robust region-growing based segmentation algorithm [10]: The skin colour segmentation is performed on predefined thresholds in the U,V-space of the video signal. Then, a blob recognition identifies the hands and the head in the sub-sampled image. Based on this information, a region growing approach segments the complete skin-colour region of the hands and the head quite accurately (see **Fig. 2**).
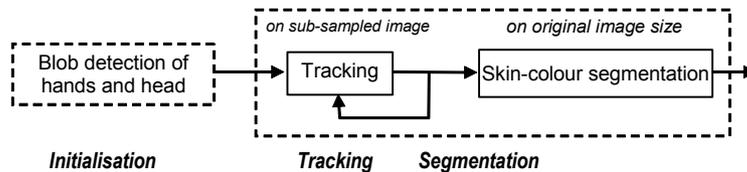

Fig. 2. Block diagram of the segmentation and tracking method of the hands

The initialisation is performed and segmentation and tracking start as soon as three separated skin-colour blobs are detected. The blobs will be assigned to hands and the head supposing that hands are initially below the head, which holds for general poses. The approach achieves real-time performance due to tracking on sub-sampled images and skin-colour segmentation limited to bounding boxes circumscribing the hands and the head. Results are presented in section 6, **Fig. 7** and **Fig. 6**.

In this context, it is a specific problem to resolve overlapping between different skin-coloured regions such as hands and the head. If a hand has contact with the face, the following process is carried out: In addition to the hands, the head blob of the participant resulting from the initialisation phase is tracked as well in the sub-sampled image, using a third bounding box. If one of the hand boxes overlaps with the head box, then only the non-overlapping part of the hand box is considered for tracking the centre of gravity. More details can be found in [10].

4   Facial feature extraction

The aim of facial feature tracking is to obtain sufficient information in order to derive a convincing and reliable rotation of the operator's head. As a result from the segmentation and tracking algorithm described in the previous section, the bounding box of the operator's head is used as a starting point for facial feature extraction. The skin-coloured pixels inside the bounding box are marked and a standard feature tracker is applied to this limited face region. The feature tracker is based on a two-step approach. First, relevant features are selected by using corner operators such as Moravec or Harris detectors. Secondly, the selected features are then tracked continuously from frame to frame by using a feature 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 distorted due to perspective changes or even become invisible and get lost. In **Fig. 3**, markers of selected features are shown in the face region in three succeeding frames. The big cross assigns the median value of all skin coloured 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 recognized as well as skin.

Fig. 3. Facial feature tracking result of three succeeding frames

5 Reconstruction of head orientation and hand positions

The main goal of the application from section 2 is to animate an avatar by human body motion captured from real-time video. Hence, accuracy in terms of correct 3D positions of the hands or precise nick and turn angles of the head are not required. However, reliable, convincing and smooth motions are important in order to support natural representation of a dynamic virtual human. On one hand, this fact facilitates the estimation of animation parameters in some way and can be exploited for simplifications. On other hand, the extracted parameters have to be filtered and outliers must be discarded in order to provide smooth motion.

The head orientation is derived from the results provided by a facial feature tracker. Based on a few robustly tracked facial features, the head orientation can be analysed by comparing the relative motion of facial features to the projected 2D motion of the head. This 2D motion is calculated by the mean change of position of all face pixels in succeeding frames. The task is to distinguish between head rotation and pure translation. In the case of a pure translation, the relative motion of each feature compared to the motion of the mean of all face pixel positions should be zero. Just the opposite holds in the case of the rotation. In this case, the motion of the mean of all face pixel positions should be 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 with $(\overline{m}_t, \overline{n}_t)$, 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 Equ. 1 and the change of horizontal and vertical rotation is approximated by Equ.2. A scale factor γ is introduced to adopt the pixel unit to angle.

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

$$\Delta\varphi_u = \sin\left(\gamma \cdot \Delta u \cdot \tfrac{\pi}{180}\right), \qquad \Delta\varphi_v = \sin\left(\gamma \cdot \Delta v \cdot \tfrac{\pi}{180}\right). \tag{2}$$

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 less than 1. As the central viewing direction is the most relevant, the animated head will adopt to this position after a while.

The positions of the hands are derived from the result of the skin-colour segmentation and tracking module. 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 two degrees of freedom of the hand position are available, just a simplified motion model can be implemented in this case. Therefore the system is based on the assumption, that the hands of the animated 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 joint and the proportions between the upper arm and the forearm, the position of the avatar's hands can be computed from these 2D tracking results. Nevertheless, a 2D to 3D reprojection is necessary, which requires some knowledge

about the imaging process. The general projection equation for a 3D point M_w in world coordinates into a 2D point $\mathbf{m}$ in image coordinates is as follows:

$$s\widetilde{\mathbf{m}} = \mathbf{P}\widetilde{M}_w \quad \text{with} \quad \mathbf{P} = \begin{bmatrix} \mathbf{p}_1^T & p_{14} \\ \mathbf{p}_2^T & p_{24} \\ \mathbf{p}_3^T & p_{34} \end{bmatrix} = \begin{bmatrix} a_x\mathbf{r}_1^T + u_0\mathbf{r}_3^T & a_x t_x + u_0 t_z \\ a_y\mathbf{r}_2^T + v_0\mathbf{r}_3^T & a_y t_y + v_0 t_z \\ \mathbf{r}_3^T & t_z \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} \mathbf{r}_1^T \\ \mathbf{r}_2^T \\ \mathbf{r}_3^T \end{bmatrix} \tag{3}$$

The matrix $\mathbf{P}$ is called the general projection matrix containing the internal camera parameters (u_0, v_0, a_u, a_v) and the external parameters ($\mathbf{R}$, $\mathbf{t}=(t_x,\ t_y,\ t_z)$) relating the world coordinate system to the camera coordinate system. For both components of the 2D image point, we get the following two reconstruction equations:

$$(1)\ \ (\mathbf{p}_1 - u\,\mathbf{p}_3)^T M_w + p_{14} - u\,p_{34} = 0, \qquad (2)\ \ (\mathbf{p}_2 - v\,\mathbf{p}_3)^T M_w + p_{24} - v\,p_{34} = 0 \tag{4}$$

These two equations have in general three unknowns, the three components of the 3D point, which can only be solved, if a second image of another camera is available, the so-called stereo case. As mentioned previously, it is assumed, that the hand position is fixed in a predetermined depth plane, which results in a fixed and known Z_w coordinate. In this case, a reconstruction becomes possible. Furthermore, we are able to avoid precise calibration as just a general transformation between the 2D image plane and the 3D plane at fixed depth is required. The setup of the camera related to the world coordinate system is shown in **Fig. 4**. If we assume just a horizontal rotation of the camera, a translational shift in y- and z-direction and a coinciding origin of the image coordinate system with the principal point, we get the following simplified projection matrix (see Equ. 5). The reconstruction equation (Equ. 4) for the desired X_w and Y_w coordinates becomes quite simple as shown in Equ. 6.

$$\mathbf{P'} = \begin{bmatrix} a_x & 0 & 0 & 0 \\ 0 & a_y r_{22} & a_y r_{23} & a_y t_y \\ 0 & r_{32} & r_{33} & t_z \end{bmatrix}, \quad \text{with} \quad \mathbf{R} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & r_{22} & r_{23} \\ 0 & r_{32} & r_{33} \end{bmatrix}, \quad u_0 = v_0 = t_x = 0 \tag{5}$$

$$a_1 X_W + a_2 Y_W + a_3 = 0 \qquad \text{with} \quad a_1 = a_u, \quad a_2 = -u\,a_v\,r_{23}, \quad a_3 = -(u\,r_{33}\,Z_w + u\,t_z)$$
$$b_1 Y_W + b_2 = 0 \qquad \text{with} \quad b_1 = a_v\,r_{22} - v\,a_v\,r_{23}, \quad b_2 = (r_{32} - v\,r_{33})Z_w + a_v\,t_y - v\,t_z \tag{6}$$

The resulting reconstructed point in world coordinates is finally:

$$X_W = a_2 b_2 - b_1 a_3 / a_1 b_1, \qquad Y_W = -b_2 / b_1 \tag{7}$$
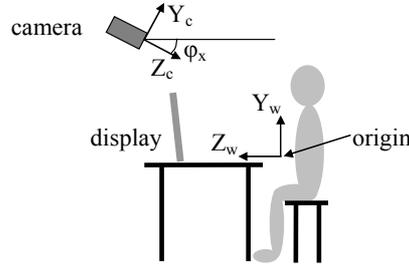
**Fig. 4.** Simplified camera setup of the considered scenario

The horizontal rotation and the translation of the camera have been measured approximately, whereas the horizontal and vertical scale factor have been chosen experimentally.

6   Prototype results

The presented approach is fully integrated in a real-time application including transmission of animation parameters and display of the animated avatar. In the following figures, several snap shots of a video sequence are presented, which show the actual pose of the user and the animated pose of the virtual character. In the right part of each example, the box around the head and the hand boxes including the segmented skin-colour pixels are shown. In **Fig. 5** and **Fig. 6** (right), the transformed head rotation is demonstrated resulting from the 2D video images of the operator. In **Fig. 6** (left), a sequence of head nick and turn angles is presented. The sinusoidal behaviour resulting from up and down (left and right) motion, is clearly visible. Interestingly, the nick angle changes since horizontal head turn was performed. The reason is caused by the mounting of the camera, which is looking from 0.4m above the head with an angle of 30 degrees. In **Fig. 7**, the animation of the avatar by moving hands is shown.

Fig. 5. Example snap shots for a head turn
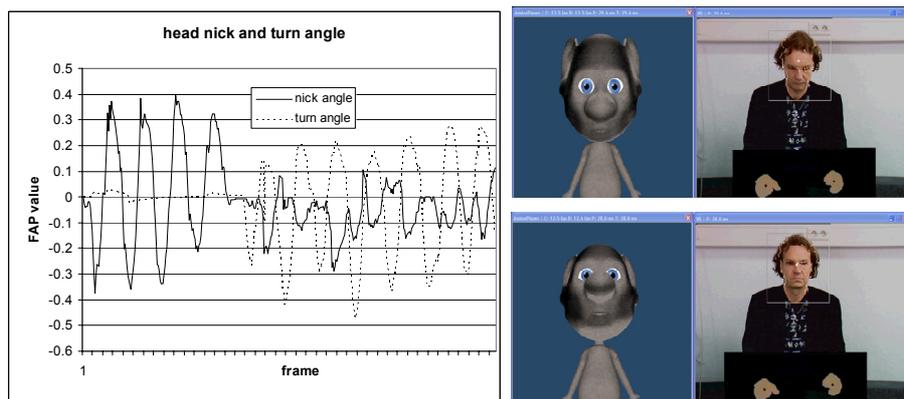
Fig. 6. Sequence of registered nick and turn angles described as FAP value (left), example images for nicking the head (right)

Fig. 7. Example snap shots for moving hands

6   Conclusion

In this paper, we have presented a complete system which uses several modules in order to steer an avatar based on live motion of a person captured by a single camera. The approach is running in real-time on a standard PC. The algorithms are robust in terms of different users, arbitrary gestures and with regard to the initialisation of the complete tracking and segmentation system. An automatic initialisation prevents the user from difficult setup procedures or specific initial gestures. This is particularly important in consumer applications, where user friendliness and easy usage play a significant role.

References

1. T.B. Moeslund and E. Granum (2001) A survey of computer vision-based human motion capture, Computer Vision and Image Understanding, vol. 81, no. 3, 231-268.
2. T. Jaeggli, T.P. Koninckx and L. Van Gool (2005) Model-based Sparse 3D Reconstructions for Online Body Tracking. Proceedings of IS&T/SPIE's 17th Annual Symposium on Electronic Imaging - Videometrics VIII, vol.5665, San Jose, California, USA.
3. I. Pavlovic, R. Sharma, T.S. Huang (1997) Visual interpretation of hand gestures for human–computer interaction: a review. IEEE Trans. on PAMI, 19: 677–695.
4. A. Just, S. Marcel and O. Bernier (2004) HMM and IOHMM for the recognition of Mono- and Bi-manual 3D Hand Gestures. British Machine Vision Conf., Kingston Univ. London.
5. P. Eisert and B. Girod (1998) Analyzing Facial Expressions for Virtual Conferencing. IEEE Computer Graphics & Applications: Special Issue: Computer Animation for Virtual Humans, vol. 18, no. 5, pp. 70-78.
6. T. Horprasert, Y. Yacoob, L.S. Davis (1996) Computing 3-D head orientation from a monocular image sequence. 2nd Int. Conf. on Automatic Face and Gesture Recogn. , p.242.
7. Z. Zhu, Q. Ji (2004) 3D Face Pose Tracking from an Uncalibrated Monocular Camera. Int. Conf. on Pattern Recognition, Workshop on Face Processing in Video, Washington DC.
8. R. R. Anderson, J. Hu, and J. A. Parrish (1981) Optical radiation transfer in the human skin and applications in in vivo remittance spectroscopy. In R. Marks and P. A. Payne, editors, Bioengineering and the Skin, MTP Press Limited, chapt. 28, pp. 253-265.
9. M. Störring, H.J. Andersen, E. Granum, (1999) Skin colour detection under changing lighting conditions. Symp. on Intelligent Robotics Systems, pp. 187-195.
10. S. Askar, Y. Kondratyuk, K. Elazouzi, P. Kauff, O. Schreer (2004) Vision-based Skin-Colour Segmentation of Moving Hands for Real-Time Applications. Proc. of 1st European Conference on Visual Media Production (CVMP), London, United Kingdom.