

# TOWARDS ROBUST INTUITIVE VISION-BASED USER INTERFACES

*Oliver Schreer<sup>1</sup>, Peter Eisert<sup>1</sup>, Peter Kauff<sup>1</sup>, Ralf Tanger<sup>1</sup>, Roman Englert<sup>2</sup>*

<sup>1</sup>Fraunhofer Institute for Telecommunications/Heinrich-Hertz-Institut,  
Einsteinufer 37, 10587 Berlin, Germany

{Oliver.Schreer, Ralf.Tanger, Peter.Eisert, Peter.Kauff}@fraunhofer.hhi.de

<sup>2</sup>Deutsche Telekom Laboratories, Ernst-Reuter-Platz 7, 10587 Berlin, Germany  
roman.englert@telekom.de

## ABSTRACT

In future videocommunication services, the user's communication device, such as PC, laptop, PDA or mobile phone is equipped with new interaction modalities. These can be cameras and microphones on the capturing side and speech synthesis and video/3D graphics on the rendering side. Haptic and tactile interfaces become also available. These modalities help the user to interact more intuitive with complex devices and tools and provide new services. Hence, a key challenge of new modalities is robustness and stability under general conditions in arbitrary environments. Furthermore, inexperienced users should be able to use these new capabilities without dedicated knowledge of device settings or algorithms. In this paper, we will present some key components for a robust vision-based user interface, which are integrated in an advanced future videocommunication service.

## 1. INTRODUCTION

Video analysis leads to many new applications such as video surveillance, person identification, motion capture for entertainment industry or medical purposes but also tools and systems, where the user can interface a device e.g. gesture recognition, human-machine interaction or interactive web-based commercial applications. One key challenge is robustness under general working conditions. In addition, the user causes many challenges like different knowledge about the system and its operation or unexpected behavior. 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. In this paper, we will present some example approaches, which are successfully integrated in a vision-based chat application. This system is the result of a joint cooperation with Deutsche Telekom Laboratories and France Telecom R&D. The main task is to capture and track the user's motion with a single camera and then to animate an avatar on the receiver side. In comparison to current chat services, a new kind of video representation of the remote

chat partner is provided without transmitting video data.

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]. Hand gesture recognition is reviewed in [2] and a 3D gesture recognition system is presented in [3]. 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 [4]. In [5], some specific face features are tracked in order to recover the orientation and position of the users 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 the next section, some basic challenges of this real-time application are listed. Then, the core algorithm for skin-color based segmentation and tracking is explained with some details on automatic initialization of the system. In the next section, the facial feature tracking component is presented, which allows a robust user independent estimation of the head rotation. The third module is a gesture recognition algorithm, which is used to animate the avatar with realistic gestures. Finally, an outlook is given regarding automatic color segmentation parameter estimation.

## 2. CHALLENGES OF VISION-BASED USER INTERFACES

Based on an example application, an advanced vision-based chat system, we will depict some general challenges, which occur in vision-based user interfaces. The block diagram of the application is shown in Fig. 1. The captured video on the sender side is analyzed by the following modules: skin-color segmentation and tracking, facial feature point

tracking and gesture recognition. Based on the video information, the position and gesture of the hands and the orientation of the head are recognized and converted to facial animation parameters (FAP) and body animation parameters (BAP) as standardized in MPEG-4 (Part 2 (Visual)). In Fig. 2, the original video and the animated avatar are shown.

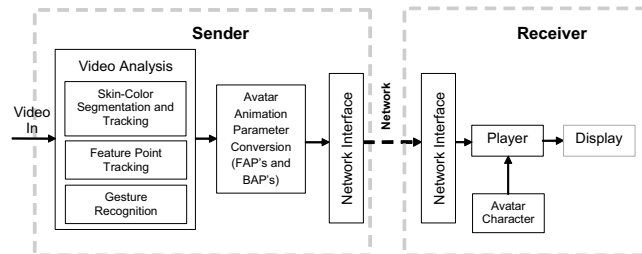


Fig. 1. Block diagram of the vision part of the chat communication system using an animated avatar

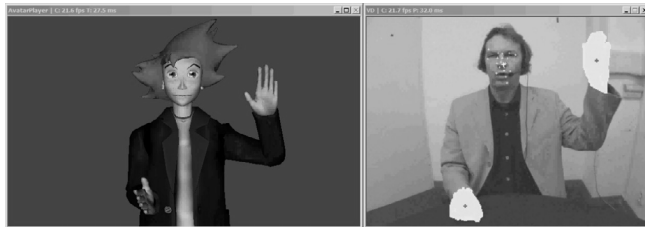


Fig. 2. Original video (right) and animated avatar (left) (the avatar is provided by France Telecom R&D)

The current system includes voice capturing and transmission. 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. For further details refer to [6].

In terms of robustness and user-friendliness the following challenges turn out:

**1. Automatic and imperceptible initialization of the tracking:** The user may not perform any initial gestures or specific positions in order to start the processing. The user must be allowed to behave natural, while using the system. The initial assignment of skin color blobs to left/right hand and head must be correct.

**2. Robust and stable tracking of hands and head:** The tracking of hands must perform robust under all conditions, specifically overlap of hands as well as overlap of hands with the face must be considered and processed. The surrounding environment should not underlay specific constraints in terms of the background, moving objects or lightning conditions. Furthermore, the algorithm must be user independent.

**3. Correct gesture recognition:** Some predefined gestures should be recognized immediately, whereas false detection must be avoided.

In the next three sections, we present some modules, which cover these challenges. All of them are integrated in a real-time chat videocommunication system, that is currently installed in a show room of the headquarter of Deutsche Telekom in Bonn, Germany.

### 3. AUTOMATIC INITIALIZATION AND TRACKING

The color of human skin is a striking feature to track and to robustly segment the user's hands and face. Hence, the human skin-color can be defined as a "global skin-color cloud" in the color space [7]. This is utilized successfully in a fast and robust region-growing based segmentation algorithm [8]. The skin color segmentation is performed on predefined thresholds in the U,V-space of the video signal. In addition to the base algorithm, the luminance channel is also used as in nearly black or very bright regions. Skin-color may be detected, although there is obviously no skin-colored region. The segmentation in the luminance channel improves the overall stability.

The algorithm starts with a blob recognition, which identifies the hands and the head in the sub-sampled image. Based on this information, a region growing approach segments the complete skin-color region of the hands and the head quite accurately (see Fig. 2).

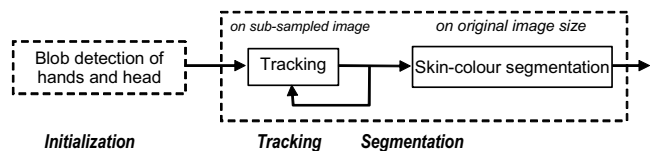


Fig. 3. Block diagram of segmentation and tracking

The initialization is performed and segmentation and tracking are started as soon as three separated skin-color 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 key idea for blob-detection is a quick evaluation of row and column histograms of the segmented image. The image is divided in three equal stripes in horizontal and vertical direction. For each stripe the maximum in the corresponding histogram interval is determined. The points of intersection of the horizontal and vertical maxima yield nine potential positions of the centers of gravity of possible hand or head blobs (Fig. 4, left). Obviously some of them have to be wrong. A neighborhood analysis searching for the points with the most skin-colored neighbor pixels removes wrong points. The resulting three positions mark left and right hand and face (Fig. 4, right). If the hands get lost during tracking, the head position is exploited to re-initialize immediately the system by evaluating the row and column histogram omitting the head area. If one of the hand boxes overlap with the head box, then only the non-overlapping part of the hand box is considered for tracking the center of

gravity. More details can be found in [8]. The approach achieves real-time performance due to tracking on sub-sampled images and skin-color segmentation limited to bounding boxes circumscribing the hands and the head.

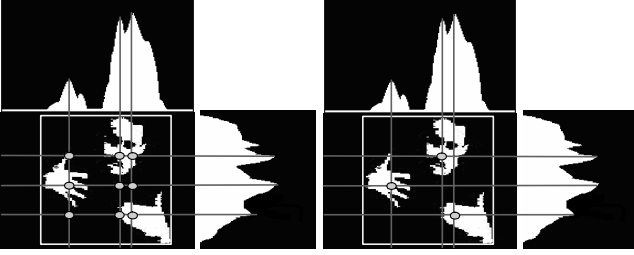


Fig. 4. Initial blob-detection by histogram analysis

#### 4. FACIAL FEATURE EXTRACTION

The aim of facial feature tracking is to derive a convincing and reliable rotation of the user's head. As a result from the segmentation and tracking algorithm, the bounding box of the user's head is used for facial feature extraction. The skin-colored pixels inside the bounding box are marked and a standard feature tracker is applied to this bounded image region. In contrast to other approaches, a general feature tracker is applied instead of a template matcher for specific face templates. 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. 5, the features are shown in the face region in three successive frames. The big cross assigns the center of all skin pixels. The considered skin color 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. Based on a few 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. This 2D motion is calculated by the mean change of position of all face pixels in successive frames. The task is to distinguish between head rotation and pure translation. In the case of a pure



Fig. 5. Facial feature tracking result of three successive frames

translation, the relative motion of each feature compared to the motion of the mean of all face pixel positions should be close to 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 behavior 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  $(\bar{m}_i, \bar{n}_i)$ , whereas the mean of all face pixel positions is denoted by  $(\bar{p}_i, \bar{q}_i)$ . 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  $\gamma$  is introduced to adopt the pixel unit to an angle.

$$\Delta u = (\bar{m}_i - \bar{m}_{i-1}) - (\bar{p}_i - \bar{p}_{i-1}), \quad \Delta v = (\bar{n}_i - \bar{n}_{i-1}) - (\bar{q}_i - \bar{q}_{i-1}) \quad (1)$$

$$\Delta \phi_u = \sin(\gamma \cdot \Delta u), \quad \Delta \phi_v = \sin(\gamma \cdot \Delta v). \quad (2)$$

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, the animated head will adopt to this position after awhile.

#### 5. GESTURE RECOGNITION

The skin-color based segmentation algorithm provides a quite accurate silhouette of the user's hand. Even in the case of frontal position of the hand to the camera some typical gestures can be evaluated. If these gestures are recognized, they can immediately be transferred to the avatar on the receiving side. The key challenge in this task is twofold: Gestures, shown by the user must be identified correctly in a very short time. On 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 recognition of true positive events. This must hold for the whole session.

In order to reduce the false negatives, the gesture recognition starts, if the following two conditions are fulfilled:

1. Spatial condition: The user's gesture can be expected in a certain region of the image, which can be defined in advance due to the viewing space of the camera.
2. Temporal condition: A specific gesture is in general performed without moving the hand. Hence, the motion of the hand can be exploited and must be small in the case of a shown gesture.

If the two conditions are fulfilled, a gesture shown by the user can be assumed. Based on the accurate silhouette of the hand, its contour is analyzed in order to derive some typical

gestures. This approach has been applied successfully in other systems (e.g. [9]). The main idea is to detect the number of fingers and their orientation. These parameters provide sufficient information in order to recognize the OK gesture, the victory sign and many more. For each contour point, the distance from its tangent along the contour to the opposite side of the contour is calculated. In the resulting distance function, the peaks assign clearly the fingers in the silhouette. Analyzing the number of peaks and the orientation of the distance vector in the silhouette provides the resulting gesture. In Fig. 6, the silhouette of a victory gesture is shown. The black lines assign the largest distances in the contour function, which coincide with the orientation of the two fingers. In Fig. 7, two screen shots of recognized gestures are shown.

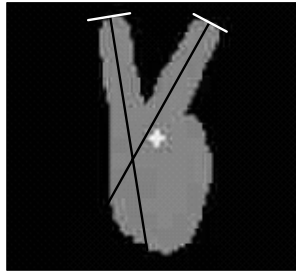


Fig. 6. Silhouette of the victory gesture and marked distances

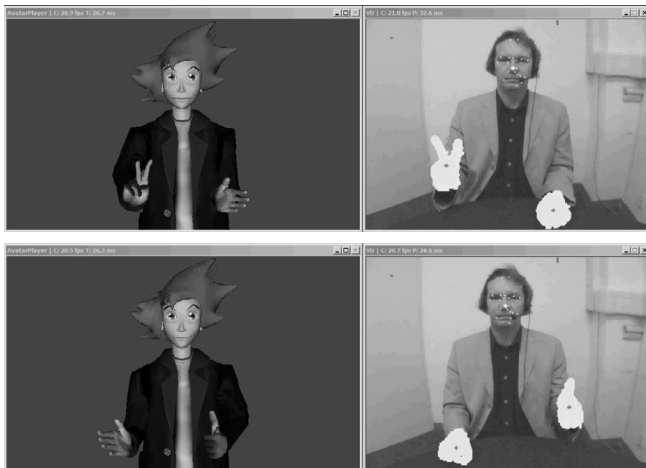


Fig. 7. Example screen shots showing two different gestures

## 6. FUTURE WORK

A crucial problem is the choice of correct skin color segmentation parameters. Manual selection can be done online, but it requires some knowledge, which cannot be assumed in general. Hence, an automatic parameter selection is planned in an initial phase of system startup. A combined motion detection and connected region selection will be implemented in order to detect the user and its typical skin color range. Due to this approach, the system becomes additionally robust in terms of changing lightning conditions, different users and camera settings.

## 7. CONCLUSION

In this paper, we discussed the challenges of robust vision-based user interfaces, which become relevant in the next

years. Interfacing a system with computer vision leads to many new applications. In the presented example system, the user can animate an avatar based on live motion. Several issues in terms of robustness and user-friendliness have been considered in different modules of the system, which are skin-color segmentation and tracking, facial feature tracking and gesture recognition. The algorithms are robust in terms of different users, gestures and with regard to the initialization of the complete tracking and segmentation system. An automatic initialization 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. The presented system is continuously demonstrated in the show room at the head quarter of Deutsche Telekom in Bonn.

## 8. ACKNOWLEDGEMENT

We gratefully thank France Telecom R&D, for the provision of the avatar. The work is funded by Deutsche Telekom Laboratories, Germany.

## 9. REFERENCES

- [1] T.B. Moeslund and E. Granum "A survey of computer vision-based human motion capture", *Computer Vision and Image Understanding*, vol. 81, no. 3, 231-268, 2001.
- [2] I. Pavlovic, R. Sharma, and T.S. Huang "Visual interpretation of hand gestures for human-computer interaction: a review", *IEEE Trans. on PAMI*, 19: 677-695, 1997.
- [3] A. Just, S. Marcel and O. Bernier, "HMM and IOHMM for the recognition of Mono- and Bi-manual 3D Hand Gestures", *British Machine Vision Conf.*, Kingston Univ. London, 2004.
- [4] P. Eisert and 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.
- [5] T. Horprasert, Y. Yacoob, L.S. Davis, "Computing 3-D head orientation from a monocular image sequence" *2nd Int. Conf. on Automatic Face and Gesture Recogn.*, p.242, 1996.
- [6] T. Ezzat, T. Poggio, "Miketalk: A talking facial display based on morphing visemes," Proc. of IEEE Computer Animation, Philadelphia PA, USA, 8-10 June, 1998, pp. 96-102.
- [7] M. Störring, H.J. Andersen and E. Granum, "Skin colour detection under changing lighting conditions" *Symp. on Intelligent Robotics Systems*, pp. 187-195, 1999.
- [8] S. Askar, Y. Kondratyuk, K. Elazouzi, P. Kauff and 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.
- [9] 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, Cagliari, Italy, 2005.