Towards Sociable Virtual Humans: Multimodal Recognition of Human Input and Behavior

Christian Eckes, Konstantin Biatov, Frank Hülsken, Joachim Köhler, Pia Breuer, Pedro Branco and L. Miguel Encarnacao

Abstract—**One of the biggest obstacles for constructing effective sociable virtual humans lies in the failure of machines to recognize the desires, feelings and intentions of the human user. Virtual humans lack the ability to fully understand and decode the communication signals human users emit when communicating with each other. This article describes our research in overcoming this problem by developing senses for the virtual humans which enables them to hear and understand human speech, localize the human user in front of the display system, recognize hand postures and to recognize the emotional state of the human user by classifying facial expression. We report on the methods needed to perform these tasks in real-time and conclude with an outlook on promising research issues of the future.**

Index Terms—**Man-Machine communication, avatars, gesture recognition, speech recognition, affective computing.**

I. INTRODUCTION

One of the biggest obstacles to natural interaction with virtual humans is their lack of understanding of user's intentions and desires. The virtual character implemented in Microsoft's Office product, e.g. the famous paper clip, illustrates this dilemma: The computer tries to infer the desires of the human user based on limited knowledge and-often incorrect-assumptions, which consequently, despite "best intentions" produce only frustration and annoyance. The result is well known: the virtual character is turned off by the human user and is cast to exile. This example illustrates that the intentions of the users are often unknown to the system and it must be avoided at any costs to act inappropriately. Hence, we must try to infer the intentions and the internal state of the user even when such a task may be never completely solved. However, this is forgivable since even humans fail in this task sometimes.

In the following, we report on methods to recognize human speech, to locate humans in front of the display system and to identify their hand posture and facial expressions.

II. AUTOMATIC SPEECH RECOGNITION

2.1 Introduction

Automatic speech recognition (ASR) is an important part for any system interacting with the human user. Virtual humans must react to utterance of the users robustly and instantaneously. The system must understand what the user says in order to generate appropriate actions and responses of the virtual humans. This problem is usually divided into two different tasks: The first one aims at recognizing robustly the sentences the user has uttered towards the avatars by generating textural transcripts or word lattices out of the captured audio signals. The second problem lies in trying to understand the language of the human user has used to communicating with the system. The latter problem falls in the domain of natural language understanding. Both tasks must still be considered unsolved, though there are already many well-working solutions in various applications and domains.

We now present the requirements and discuss the approach we followed to support natural communication with virtual humans.

2.2 Requirements for ASR in Virtual Human

Generically ASR translates spoken language into a suitable representation computers can work more easily with. Words and sentences must be recognized using the audio signal of the speech. ASR should either produce the most likely transcription, a ranked list of n-best results, or word lattices which may be analyzed later on by using techniques based on natural language understanding.

Besides these rather general requirements, additional constraints posed by this project were:

- Real-time processing with low latency: The system must run online with small latency since any delay between the utterance of the human user and the response of the avatars w break the direct communication between the participants.
- Speaker independence: The recognition must be independent from the speaker. As the potential users are male and female children and adults, quite general acoustic

Manuscript Received on February 21, 2007.

Christian Eckes received the diploma degree in Physics from the University of Dortmund in 1995. E-Mail: Christian.Eckes@iais.fraunhofer.de.

Konstantin Biatov received master degree in applied mathematics from Moscow Institute Physics and Technology, Moscow, Russia in 1973 and Ph.D degree from Glushkov Institute of Cybernetics NAS Ukraine, Kiev in 1986. E-Mail: kostyantyn.biatov@iais.fraunhofer.de .

Frank Hülsken is a mathematician with background in pattern recognition. His research interests include speech regognition, image processing, 3D reconstruction and animation. E-Mail: frank.huelsken@iais.fraunhofer.de.

Joachim Köhler received his diploma and Dr.-Ing. degree in Communication Engineering from the RWTH Aachen and Munich University of Technology in 1992 and 2000, respectively.

Pia Breuer Currently works as a research assistant at the research group Media Systems at the University of Siegen.

Pedro Branco graduated in Computer Science from University of Porto in 1997 and received the doctorate degree in Information Systems from University of Minho.

L. Miguel Encarnação is the Chief Technology Officer and Executive Vice President of the IMEDIA Inc., a Rhode Island, U. S. private research lab.

models must be trained. Ad-Hoc adaptation may be used but without any explicit learning phase.

- Multi-language: English and German language must be supported to ensure demonstrations of the project in an international environment.
- Adjustable language models: Prior knowledge about the context of the dialog and expected answers may be used as grammars to speed up the recognition.
- Distributed architecture supporting multiple users: Since all components consume significant processing power and may also need special and potentially incompatible hard- and software, a distributed client-server architecture based on Ethernet communication has been chosen for integration. Moreover, at least two persons should be able to use the speech recognition system simultaneously.
- Concurrent operation: The system must react flexible to the utterances of the human users by continuously monitoring and classifying the audio signal. The user must not need to switch the ASR explicitly on and off, as it is common practice when using walky-talkies.

2.3 VoiceXML-Based ASR architecture

Test and evaluations of concepts and ideas within an integrated research project can only be achieved if the different modules are integrated as early as possible. Therefore, we have adapted a commercial dialog system VoiceGenie from VoiceGenie Technologies (now part of Genesys) based on VoiceXML and Voice-Over-IP (VoIP) technology in order to bootstrap the multimodal recognition model in the course of the development of the Early Demonstrator in the first 6 months of the project. A microphone was connected to a voice server module (SIP) which captures the audio signal and transports the signal via VoIP to the VoiceGenie server. A Java client running on a Tomcat JSP server controls the VoiceGenie server and defines the language and dialog structure of VoiceGenie based on VoiceXML documents. The client program is also connected to the dialog engine to adapt the current language model to the current situation and to inform the dialog model about the recognized sentence which is encoded as an XML document. The user acts in the role of a student answering the questions during the lesson by verbal communication instead of using the mouse.

The Voice Genie solution within the Early Demonstrator has proven quite reliable. First concepts of language models, dialog structures and experiments of the interaction scenarios could be evaluated (e.g. see, for instance, the evaluation of the Early Demonstrator in the article “Business Cases for Virtual Human Technology: Evaluation and Exploitation” in this publication). However, due to the fixed and propriety architecture we chose to develop our own speech recognition system based on open source toolkits.

2.4 German ASR with open source speech recognition toolkit ISIP

Our development of the German spoken dialog system in the Virtual Human project is based on the ISIP toolkit [15]. This is an open source toolkit for developing large vocabulary speaker-independent speech recognition applications based on the state-of-the-art Hidden Markov Model (HMM) technology for ASR. The toolkit includes training, decoding and evaluation modules. The training module can train acoustic model with the different complexity such as monophone, word-internal triphone or cross-word triphone based acoustic models. ISIP supports decoding based on a language model presented as a probabilistic finite state machine (grammar) or as an n-gram language model.

For grammar processing ISIP provides various grammar compiler tools. The output of ISIP decoder can be presented as 1-best, n-best or as a lattice. The ISIP toolkit also includes an extension specially developed for real time applications. In particular, this extension includes the possibility to switch online grammar depending on the context of the dialog without stopping ASR application.

In the described speech recognition system the input signal was a 16 kHz audio signal with 16 bit sampling. The ASR had the following components:

- speech/non-speech analyzer,
- feature extractor for speech data,
- decoder,
- output converter.

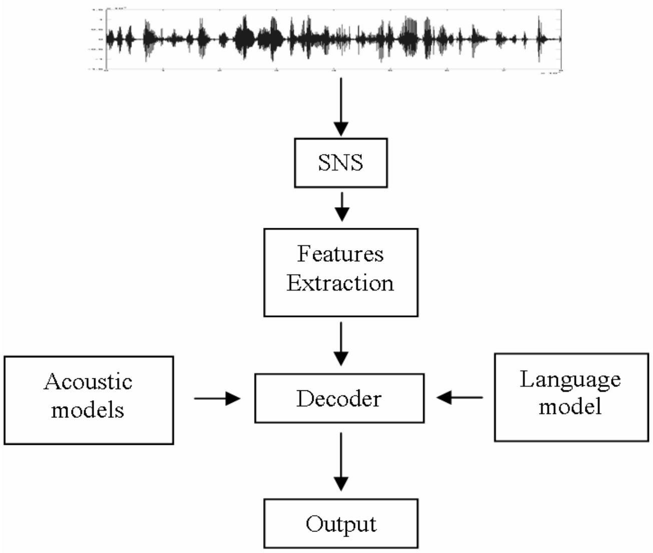


Fig. 1. Workflow in speech recognition.

The workflow of speech recognition system is presented in Fig. 1. The speech recognition application envisioned for Virtual Human is working without any on/off button or switch connected to the microphone which is usually used in order to simplify the search for the start and end positions of the captured speech signal. Hence, we have developed an automatic Speech-Nonspeech classifier module which detects speech automatically based on a silence ratio It calculates the percentage of samples from the audio stream which are less than a predefined threshold (silence ratio) for sequences of equal-size sliding windows. For each pair of neighboring windows, the ratio between their silence ratios is calculated. When it exceeds the predefined threshold the audio is considered the start of the speech. When this ratio exceeds for the last time the same threshold is considered the end of the speech. The audio signal is recorded in ring buffer in parallel with the speech/non-speech analysis.

The next module, the feature extractor, analyses the features only from the speech segments. The feature vectors consist of 12 mel-cepstral coefficients plus energy extended with delta mel-cepstral coefficients and delta energy, delta-delta mel-cepstral coefficients and delta-delta energy which are frequently used in state-of-the-art speech recognition systems.

Next, the acoustic and language model are discussed. The acoustic model was trained using data developed for the Verbmobil project. Currently this data is a part of the Bayern Archive for Speech Signal (BAS). The archive includes data collected in difference acoustic conditions. The Virtual Human project uses only the data collected with the room microphone since it is the same setup used in the virtual human demonstrator. The cross-word acoustic model was trained with 8 Gaussians for each HMM state. For the training, 3482 German spoken sentences were used with the total duration 2.5 hours.

The language model is a set of finite state machines (FSM) generating templates for allowed sentence. For each sequence of spoken words the matching FSM corresponding to that sequence is used to reduce the set of possible words to be recognizable, allowing for a more robust decoding. As the language model can be exchanged on the flight, both robust and flexible dialog modeling is supported by the system.

For decoding, speech recognition, the ISIP decoder is used which is based on a state-of-the-art Viterbi decoder for HMM.

In spoken dialog is very important to provide the understanding module and dialog manager with the information which potentially include correct recognition output. To get potentially correct answer in output it is better to use output in the form of n-best or lattice instead of 1-best output. The lattice is more compact representation then n-best. It is the reason that in Virtual Human task the decoder provides lattice output. At the end of decoding the lattice is converted in MPEG-7 representation.

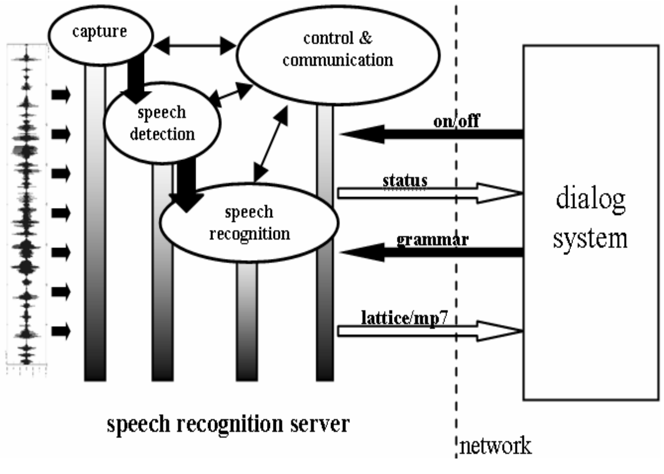


Fig. 2. System architecture of ASR in VirtualHuman.

The architecture is described in Fig. 2. As the difference tasks: audio capture, speech detection, speech recognition and network communication with a client application must be performed concurrently, the architecture consist of four processing threads, one for each sub tasks. The communication is based on shared memory with semaphore protection. The system has been implemented in ANSI C++ and works under Linux capturing with ALSA system calls.

In the second phase of the project, two human users must interact with the application which leads us to multiply this approach. We simply dispatch two different instances of the ASR server using two different virtual ALSA capture devices connected to two different TCP/IP network ports. We have used ALSA filters to route the signal of two microphones through a stereo capture jack of a standard soundcard to realize two virtual capture devices. This avoids special hardware and ensures that all project partners can evaluate the system.

III. GESTURE RECOGNITION

3.1 Introduction

Gesture recognition aims at recognizing the posture of the human body (torso, face, hands, arms) and their dynamics for various task, such as animation, training, therapeutics, gaming, security and, last but not least, human-computer interaction, as in the case of the Virtual Human. The question remains, what do we really want to recognize? After investigating the typical requirements of a virtual reality (VR) display application with virtual characters, we have chosen to focus our research on the following sub tasks of gesture recognition:

- Locate the users: The user must be located in the real world in front of the VR display. This enables virtual humans to sense the presence of the human user and to trigger appropriate behavior which makes the virtual humans, e.g. by looking towards the human user or by welcoming him or her.
- Recognize simple deistic gestures: Pointing gestures, for example, are a natural way to interact with a VR application.
- Recognize hand gestures: Simple hand gestures such as waving, thumb up, pointing with fingers and other more articulated hand postures and motions are very easy for humans to perform but difficult to be recognized automatically thereby provide excellent topics for research.

As the overall task has been defined now, let us investigate what kind of technical requirements for gesture recognition we have to fulfill within the Virtual Human project:

- Robust against “disguise”: Some display technology requires from the user to wear shutter glasses or glasses with polarization filters. The human user becomes masked and standard recognition methods, such as face detection, might not be applicable in all cases.
- Robust to illumination changes: The use of display systems, e.g. back or front projection or active tracking of shutter glasses, usually generates instable and uncontrollable illumination in the visual and/or IR spectrum. Passive

IR-technology might interfere with this type of stereo display.

- Real-time processing: The algorithms and hardware must be real-time in order to avoid strong delays and latencies in the interaction loop which may destroy the feeling of immersion and direct interaction.

Today not all of these requirements can be fulfilled simultaneously by state-of-the art display technology. Hence, two different systems and hardware setups are used within the Virtual Human project to fulfill all requirements. We focus in full-body and hand gesture recognition on detecting the human hand and full-body posture, using a special 3D-camera, which is explained in detail in the following section. On the other hand, our system for recognizing facial expression do not need such sophisticated hardware but rely on a standard (web)cam sensible to light stemming from the normal visual spectrum, instead, as we will discuss in detail in the next chapter.

3.2 Hardware: Swissranger SR-2 of CSEM

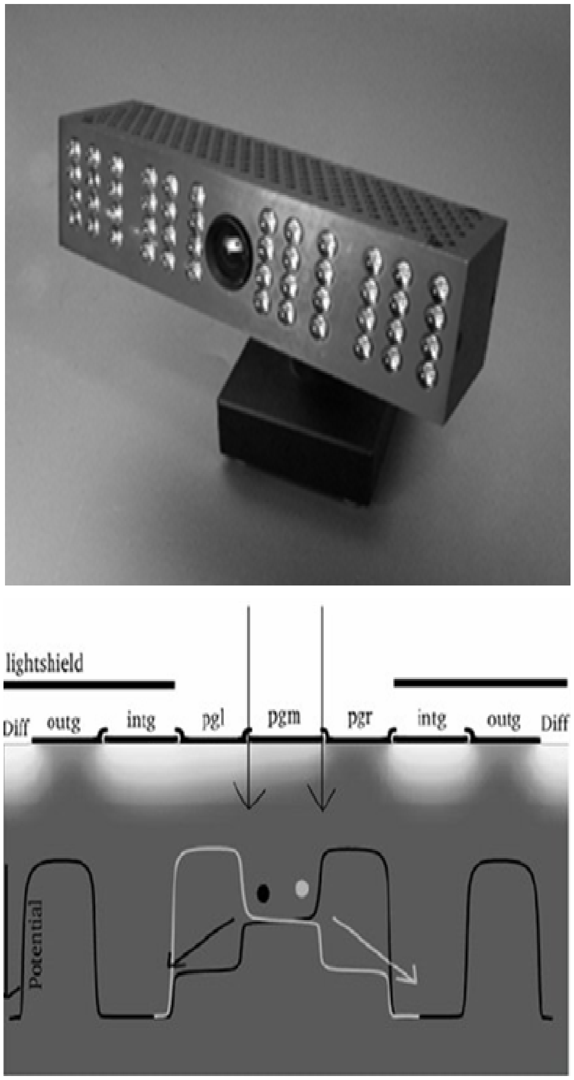


Fig. 3. The SR-2 of CSEM (up), a simulated cross section & potential distribution (down).
[taken from CSEM [12]].

We have chosen to base our work in gesture recognition on a special novel hardware which has recently been made available, the miniature infrared time-of-flight camera Swissranger SR2 developed by CSEM (see [13, 14]). See also Fig. 3. This IR camera delivers depth and intensity measurements of a size of 124x160 pixels resolution with a frame rate of up to 30 Hz. It emits modulated infrared light of 870 nm wavelength amplitude-modulated with 20 MHz. In contrast to standard IR cameras, the SR-2 does not only measure the backscattered intensity but also the depth for each pixel. The measurement is based on the Time-Of-Flight principle: Each pixel of the camera is able to measure the phase angle of the reflected IR light in relation to the current phase angle of the active modulated IR LEDs. The Fig. 3 (down part) illustrates how the demodulation of the reflected light is performed by sampling:

Each pixel contains 4 bucket integrators able to collect charges that are generated by the absorption of IR quants in the photo-sensitive semiconductor. A dynamical external electric field switches between the different integrators/capacities four times in each period. The collected charges are read out after a particular integration time has passed by the electronics. The phase shift can be computed from the 4 accumulated charges and the depth can be computed. I refer to the references for further information how this is achieved.

3.3 Localizing the Head/Hand of the human user

We have detected the human user in front of the camera by a combination of median filtering, depth keying and convolving the resulted depth image by a parameterized filter bank based on rectangle convex filters.

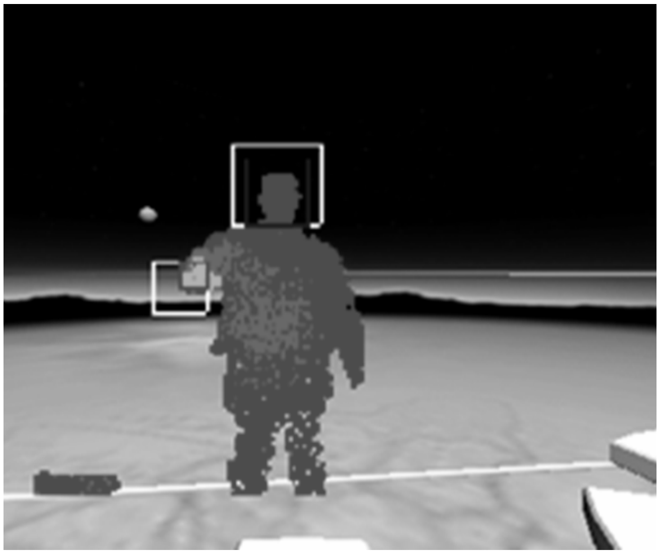


Fig. 4. Localizing the head and hand of the user

Fig. 4 shows an example of the localization of the person in front of the Early Demonstrator setting. The image depicts in the background a VR scene in which the depth camera image has been rendered as an alpha image plane. The first step in the processing pipeline is to apply a median 3x3 filter to eliminate outliners and “dead” pixels. Next, depth segmentation or keying is performed to eliminate all pixels with stems from the background. We have used a depth range of 1.5m to 2.5m from the display system; pixels with different distance are discarded. The result can be seen in Fig 4 represented by the red area. The user as well as part of a stage has been recognized by the system, transparent alpha values have been used for the discarded pixels. Connected component analysis is used to find the largest segment in the depth image.

The next step is to identify the location of the human head. We use a filter bank of simple parameterized head filters H_i which detects head-like convex structures in which the width w_i and height h_i of the human head can be parameterized while holding the aspect ratio fixed within a reasonable range. We implicitly convolve the depth image with the filter bank F_p in which each parameter set p identifies a head filter of a particular size and aspect ratio. As the kernel of the head filters can be decomposed into homogenous rectangles regions, we use an integral image to speed up the processing. The implicit filtering of the depth image results in one saliency image S_p for each filter. We compute the local maxima of these images spatially and across neighboring scale. The best matching filter is drawn in Fig. 4 around the head. The large rectangle shows the filter location, the small inner part defines a region of the filter kernel in which small depth values generate an optimal response; hence they contain constant negative values. The filter kernel between the grey and white regions is set to positive values. Hence, the head filters detect convex regions near the camera with a significant depth edge around the head.

The human hand is localized using a larger filter bank with parameterized peak detection filters and convex filters with different orientation. After finding the maximum convex and peak consistent with assumption that the hand must be nearer to the camera than the detected head, the hand position becomes localized as well. An example of a matching hand filter has been drawn around the human hand in Fig. 4.

Using a calibration of the camera, we were able to detect the position of the human head and hand in real-time, e.g. with 30 Hz. Deistic gestures can now be performed and recognized by the machine for interaction since the 3D-positions of the human head and hand are sufficient to realize a pick ray in VR systems, e.g. in Fraunhofer VRLM/H|ANIM Player Avalon or the OpenSource Render system Avango. We have recorded a corpus of 7 persons pointing towards various positions at the display system. The results are very promising and first qualitative evaluation of the precision supports already very natural applications.

3.4 Articulated Hand Recognition

We will now briefly report on the work on hand gesture recognition where we have used the Swissranger SR-2 data to identify the articulation parameters of a human hand. We followed a bottom-up approach to estimate some coarse parameters, and a model-based refinement. For more details, we must refer the reader to an upcoming publication [3].

We view hand gesture recognition as a problem composed of two sub-problems: the first is to reconstruct the hand with its articulations in real-time, the second is to match the time-depending parameters of the hand, the position of the hand, its size and all joint angles, to one of relevant hand gestures we want to recognize, e.g. a wave, clap, thump up, finger ring, a point gesture or a beat gesture. We consider the latter problem solved, or, at least, as easier since there are already well-known methods to deal with this problem, e.g. by using HMM models, as we have seen in speech recognition. The first problem is harder and therefore we have tried to focus on it in this work. Despite significant research in the field [4, 11], dynamic hand reconstruction independent from illumination remains largely unsolved.

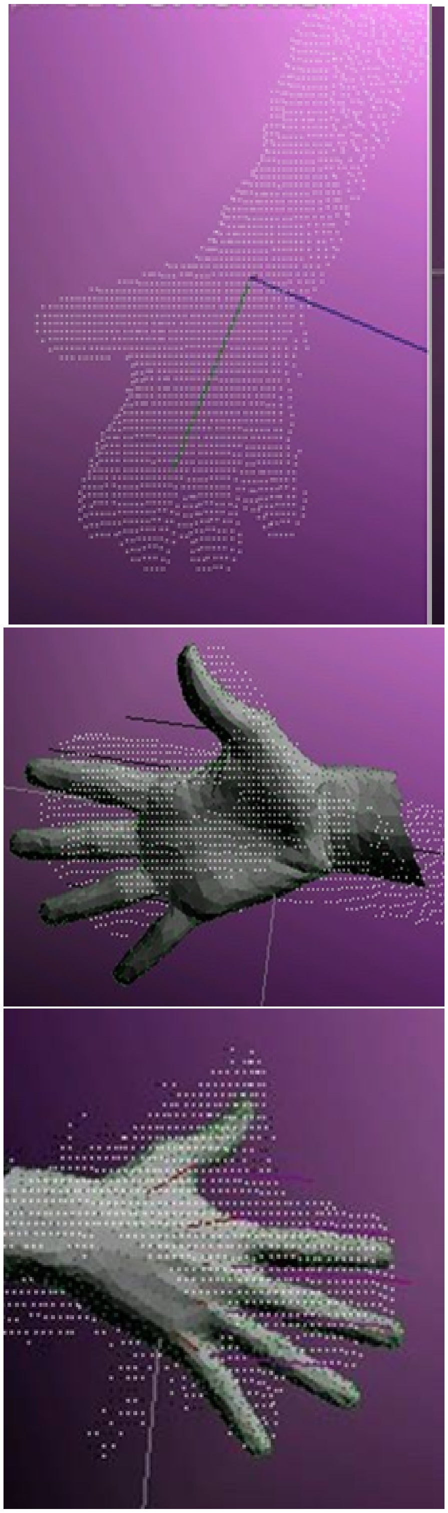


Fig. 5. Point cloud with PCA axis (top) and model-based refinement (middle, bottom)

Our steps in reconstructing the hand parameters are the following: calibration, depth keying, 3D-point cloud based PCA and a registration with an articulated hand model. We calibrate the camera by estimating the internal camera parameters (for details see [3], also [10]). The next step separates all pixels in the scene which do not belong to the human hand-arm segment. The hand is posed in front of the camera at a distance by 0.5-1 meter which defines the working volume – all pixel values with different distance are discarded. After applying a median filter with 3x3 kernel size, and depth keying, we solve the segmentation problem by performing connected component analysis and selecting the largest remaining connected sub-segment as input to the reconstruction methods. Based on our calibration, we compute a cloud of 3d-points which represent our hand and arm segment. We use heuristics based on the length of typical hand and arm to separate the hand from the arm.

A first crude estimation of the hand is obtained by fitting an ellipsoid into the data points. The center of mass and the principal axes of the point cloud provide a good estimate for fitting an articulated hand model to the point cloud, as the visualization of the results have revealed. We have used a hand model kindly provided by Irene Albrecht (see [1]) for model-based refinement. The aim of this first fine matching is to determine translation, rotation and scaling of the model accordingly. A skeleton of a three dimensional hand model is placed into the point cloud so that the distance of the surface points to the measured data is minimized. The next step in our gesture recognition systems is the fitting process which aims at minimizing the distance between model and point cloud based on the sum of point-to-point distances. In this case, the Hausdorff distance may be used as a metric. On the other hand, metrics based on pure pairwise distances between point clouds do not take benefit from the more explicit surface properties of the hand model since it "only" tries to position the point clouds together as near as possible (similar to ICP registration). Alternatively, one might compare the surfaces/visual hulls between model and reconstructed surface patches by minimizing the volume between surface patches directly since the model also consists of a triangle mesh.

The system was able to recognize 7 degrees of freedom of a human hand with 2-3 Hz frame rate without optimizing the research code. This is a promising result and defines a road map for further research. Future research should perform an evaluation of the results, increase the amount of articulation of the hand model and investigate sensor fusion, e.g. by using additional calibrated video camera to increase the still quite limited resolution of the Swissranger camera. Further work aims at improving the robustness of the system against out-liners by registering the model to the sensor point clouds with more robust algorithms, such as Random Sample Consensus (RANSAC) and to track the data by more statistical approaches [5].

IV. ANALYSIS OF FACIAL EXPRESSIONS

The next section discusses the development of a system for monitoring users facial expressions applied to the interaction with embodied virtual characters – virtual humans. We present the arguments on why within certain contexts facial expression monitoring is useful and discuss its application in a scenario of interaction with a virtual character.

4.1 Beyond the users´ voluntary responses - the importance of interlocutor behavior/body language

In social interactions, body language is a crucial aspect of the communication. The hand gestures, the body posture, the voice, the eyes, physiological manifestations such as blushing and facial expressions, are inseparable from the message, they become in part the message. The non-verbal aspects of communication should therefore be central to a system that aims at providing engaging dialogues with a virtual human; both from the point of view of the synthesis of realistic gestures, as well as the recognition of the human interlocutor's body language. For example a smile or a frown are reactions that should influence the dialogue, the flow of the storytelling, the attitudes of the virtual human, being otherwise the risk that the user perceives the interaction as cold, disconnected or just the playback of recording. This second aspect is the focus of this section.

4.2 Facial Expressions

Facial expressions deserve a special treatment when considering a multimodal conversational system. An important aspect of the interlocutors behavior is transmitted through the face. For fully capable social actors communicating within a familiar culture, the meaning and nuances of facial expression within a dialog may seem natural and intrinsic, but when dissected its complexity quickly emerges. It obeys social rules and spans through several dimensions of semantic levels. Wehrle and Kaiser enumerate the role of facial expressions according to the following categories [16]:

- In speech it is often used by the listener as a back-channel informing the speaker that he can go on talking and that he has been understood (regulator).
- To emphasize a particular message, or to change the meaning of verbal message where the speaker facial expression modifies or contradicts what is being said, e.g. when being ironic (illustrator).
- As a mean for installing, maintaining, or aborting a relationship, e.g., when a couple is discussing a controversial topic, a smile can indicate that although they disagree on the topic there is no "danger" for the relationship.
- An indicator for cognitive processes: e.g., frowning often occurs when somebody does some hard thinking while concentrated attending to a problem, or when a difficulty is encountered in a task.
- An indicator for an emotion (affect display).

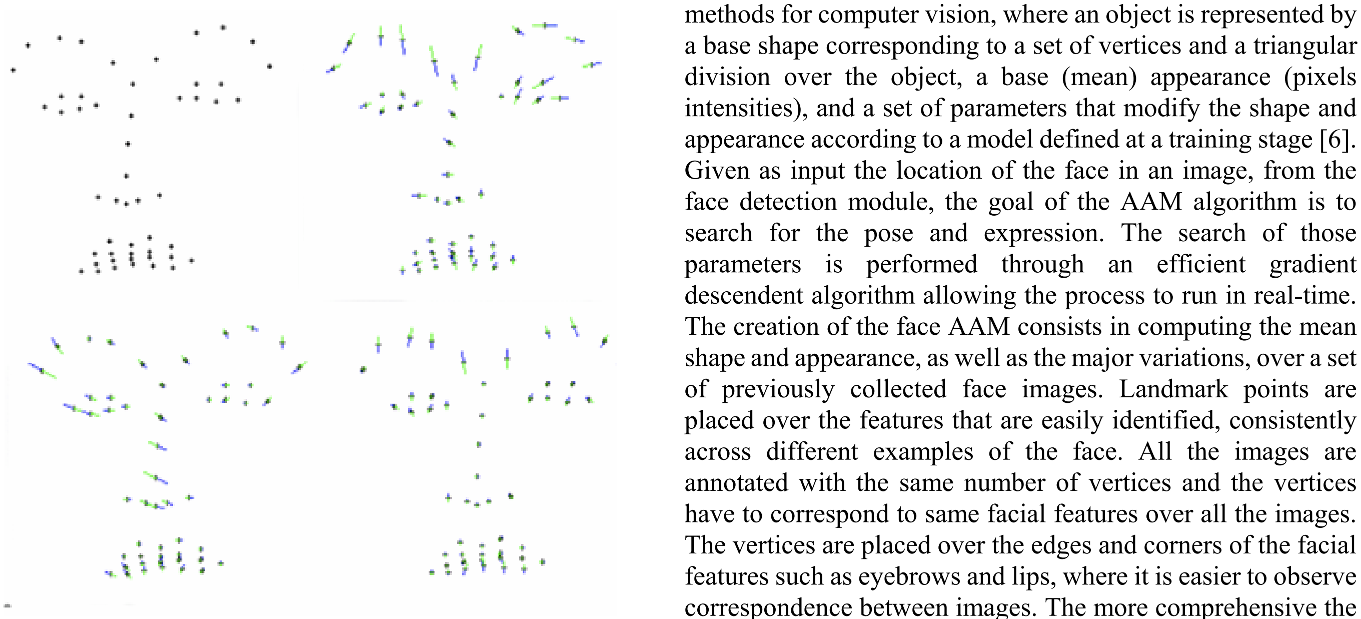
The ultimate facial expression recognition system for engaging in a realistic machine/user dialogue should be able to differentiate the subtleness of situations and understand the context. That goal is certainly too much of an ambitious endeavor at several knowledge areas. In any case the current conversational interaction systems with embodied characters are restrictive in terms of dialogues, and are, therefore, predictable to the point that at that level an important part of the subtleness of body language might be lost.

The question is opened to what extend all those levels of meaning are present in such interaction. Previous work has found that in Human-Computer Interaction (HCI) facial expressions reflect the users' difficulty with a task [8], [2], and could eventually be appropriate for a virtual character to react on that event [9]. While certainly there could be still ambiguities associates with those reactions, its integration within the dialog script of a virtual character are easier to integrate than other subtleties of human body language. Under that light and within the bounds of the virtual human project we opt for a more modest but realizable goal of identifying the interlocutor positive/negative reactions to the ongoing dialog with the virtual human.

4.3 A Video-based system for Facial Expression Analysis

Two different methods are available to record facial expressions: facial electromyography (EMG) and video-based systems. Due to the obtrusiveness of EMG sensors for generic HCI contexts, we opted for a video-based system to record and analyze the expression. We describe next a video-based facial expression analysis tool - FACEit.

FACEit is a facial expression analysis tool developed to support the research on the study of facial expressions in a HCI environment. FACEit was developed in C and is implemented on top of the Intel Open Source Computer Vision Library[2] (OpenCV) 5.0 and Intel Integrated Performance Primitives 4.1[3] (IPP). OpenCV provides a range of computer vision algorithms aimed at real-time video processing. OpenCV is made freely available by Intel. IPP provides highly optimized software functions for a variety of data processing functions. The image processing and matrix algebra are the two library components used.



(a) Base shape and shape vectors overlaid on top of the base shape

(b) Base appearance and variation

Fig. 6. Linear shape and appearance model of an independent face AAM.

FACEit is composed of three major modules: Face Detection, Facial Features Alignment and Facial Expression Analysis. The Face Detection is implemented using the module provided in the OpenCV library. It finds rectangular regions in the given image that are likely to contain frontal view faces and returns those regions as a sequence of rectangles. At the heart of the FACEit system is a fast Active Appearance Model (AAM) algorithm [12]. An AAM belongs to a group of statistical methods for computer vision, where an object is represented by a base shape corresponding to a set of vertices and a triangular division over the object, a base (mean) appearance (pixels intensities), and a set of parameters that modify the shape and appearance according to a model defined at a training stage [6]. Given as input the location of the face in an image, from the face detection module, the goal of the AAM algorithm is to search for the pose and expression. The search of those parameters is performed through an efficient gradient descendent algorithm allowing the process to run in real-time. The creation of the face AAM consists in computing the mean shape and appearance, as well as the major variations, over a set of previously collected face images. Landmark points are placed over the features that are easily identified, consistently across different examples of the face. All the images are annotated with the same number of vertices and the vertices have to correspond to same facial features over all the images. The vertices are placed over the edges and corners of the facial features such as eyebrows and lips, where it is easier to observe correspondence between images. The more comprehensive the images in the training data are in terms of identity, pose, expression and illumination, the better the AAM will be able to describe a wider variety of faces. Theoretically, a large enough

[2] www.intel.com/technology/computing/opencv/index.htm
[3] www.intel.com/cd/software/products/asmo-na/eng/perflib/ipp/index.htm

training set could account for all the shape and appearance variations. In practice, however, the larger the model the more instable the algorithm behaves [7]. In the various face AAM we trained, we limited the training dataset to the images of the specific user to monitor the expressions. This increased the robustness of our system and limited the effort of hand labeling the images to tens of images rather than hundreds or thousands. The images in the training dataset included faces under different poses and expression, since we intended the AAM to model expressions and be robust to pose variation.

After the hand labeling process, the vertices for all the images in the training set are aligned. Principal component analysis (PCA) is applied to the aligned vertices to compute the components of shape that account for the most variation in the dataset, resulting in a mean shape (s0) and shape variations {s1,...,sn}, Fig. 6a. A similar process is performed for the appearance: for each image in the training set, the pixels inside the shape defined by the hand labeled vertices, are warped back to the base shape and PCA is applied to the collection of those images to calculate the mean appearance (a0) and appearance variations {a1,...,am}, Fig. 6b.This process allows an AAM to describe new instances of the face as a sum of a mean shape and mean appearance plus a linear combination of shape and appearance variations.

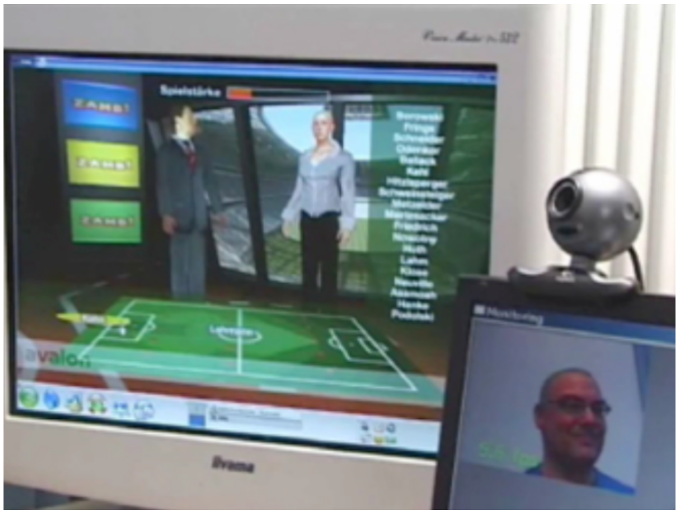
The process described previously is performed once in a training stage. The real-time process of fitting the AAM to the image is described next. Given as input the location of the face in an image, from the face detection module, the goal of the AAM algorithm is to find the best linear combination of shapes and appearances that fits the input image. In other words, the goal is to minimize the pixels intensity difference between the input image and a linear combination of {s0,..., sn} and {a0,...,am}. Different gradient descent algorithms are available to solve that expression. We implemented the method by [12] for an independent AAM. The tracking at each frame is considered successful when the difference between the input image and the modeled image falls below a given threshold. As long as the AAM fitting algorithm converges within a certain number of steps, defined by an error threshold, FACEit continues fitting the AAM to the new frames**.** When there is a failure to converge, it resets back to the face location mode.

The last step, the face expression analysis module, classifies the set of parameters found by the AAM fitting algorithm previously described as a neutral, positive or negative expression. The classification is performed using a k-nearest neighborhood approach, where the AAM fitting parameters are compared against the fitting parameters of the face images manually classified as neutral, positive or negative, used in the creation of the AAM.

4.4 Application to the VH Project

The FACEit system was tested on the virtual human project to provide real-time assessment of the user engagement to the on going dialog. On the chosen application scenario, one of the virtual humans attempts to do some humor, sometimes successfully and in other cases upsetting the user. The users' reaction triggers different paths in the storyline emphasizing the jokes or avoiding it altogether. Fig. 7 exhibits a snapshot of a positive user reaction with the associated virtual human dialogue.

While anecdotal feedback from observers and users of this particular setup on the appeal of such interaction does certainly not bear any statistical significance, related studies with virtual characters [9] show promising results that such affective monitoring might in future improve the human/virtual human dialogue. Performance, real-time synchronization, and hardware-cost challenges are issues that must be addressed though.



VH: "What does Felix Maggath (the former Bayern Munich trainer) say when Oliver Kahn runs noto the playfield?-He doesn't do anything. He just wants to play."

H: [smile]

VH: "It seems as if you are having fun. Just wait, we're have many more jokes.

Fig. 7. Tracking the expression of the user triggering the virtual human response.

V. CONCLUSIONS AND FUTURE WORK

We have shown in this article how speech, posture and emotional displays of the human user can be recognized to develop more robust and sophisticated virtual humans. Speech recognition has matured considerably but still suffers from problems due to unknown vocabulary, noise-sensitivity and limited natural language understanding. Gesture recognition is less matured and innovative methods must still be developed until virtual humans are able to understand the movements and indentions of their real alter egos. Recognition of facial expression has shown considerable progress but effective real-time monitoring still requires significant improvements. It became clear, however, that only the tight integration of those modalities will be able to overcome the ambiguities associated with any one of those forms of human-human communication. Therefore, and not surprisingly, much more research in speech recognition, human gesture recognition and effected monitoring is needed before Virtual Humans become effortlessly accepted by the human user. But the presented results and experiences gained in the Virtual Human project are encouraging and promise for significant progress to be in reach in the not so far future.

REFERENCES

[1] I. Albrecht, J. Haber and H. P. Seidel. Construction and Animation of Anatomically Based Human Hand Models, In: *SCA '03: Proceedings of the 2003 ACM SIGGRAPH/, Eurographics Symposium on Computer animation*, pp. 98–109, 2003.

[2] P. Branco, P. Firth, L. M. Encarnação and P. Bonato. Faces of Emotion in Human-Computer Interaction, In Proceedings of *the Conference on Human Factors in Computing Systems (CHI'05) Extended Abstracts*, ACM Press, pp.1236-1239.

[3] P. Breuer, C. Eckes and S. Müller. Hand Gesture Recognition with a Novel IR Time-of-Flight Range Camera -- A pilot study, Proceedings of *the MIRAGE conference "Computer Vision/Computer Graphics Collaboration Techniques and Applications"*, 2007 March, INRIA Rocquencout, France, 2007, to be published by Springer in the series LNCS.

[4] G. H. Bendels, F. Kahlesz and R. Klein. Towards The Next Generation of 3d Content Creation, In: *AVI '04: Proceedings of the working conference on advanced visual interface*, pp. 283–289, 2004.

[5] M. Bray, E. Koller-Meier and L. Van Gool. Smart Particle Filtering for 3DHand Tracking, in: Proceedings *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, Seoul, South Korea,* 17-19, pp. 675–680, 2004 May.

[6] T. F. Cootes, G. J. Edwards and C. J. Taylor. Active Appearance Models, In H. Burkhardt and B. Neumann, editors, *5th European Conference on Computer Vision*, vol. 2, pp. 484–498. Springer, Berlin, 1998.

[7] R. Gross, I. Matthews and S. Baker. Generic vs. Person Specific Active Appearance Models, *Image and Vision Computing*, vol. 23, no. 11, pp. 1080–1093, 2005.

[8] R. L. Hazelett. Measurement of User Frustration: A Biologic Approach, in Proceedings of *the Conference on Human Factors in Computing Systems (CHI '03)*, Ft. Lauderdale, Florida, USA, Extended abstracts. New York: ACM Press, pp. 734–735, April 5-10, 2003.

[9] N. Jaksic, P. Branco, P. Stephenson and L. M. Encarnação. The Effectiveness of Social Agents in Reducing User Frustration, in Proceedings of *the Conference on Human Factors in Computing Systems (CHI'06) Extended Abstracts*, pp. 917–922, 2006.

[10] T. Kahlmann and H. Ingensand. Range Imaging Sensor Properties and Calibration, in: Proceedings of *1st Range Imaging Research Day*, September 8/9, 2005 at ETH, Zurich Switzerland, pp. 71–80, 2005.

[11] X. Liu and K. Fujimura. Hand Gesture Recognition Using Depth Data, in: Proceedings of *Sixth IEEE International Conference on Automatic Face and Gesture Recognition*, Seoul, South Korea, 17-19 May 2004, *IEEE Computer Soc*, pp. 529-534, 2004.

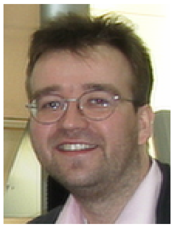
[12] I. Matthews and S. Baker. Active Appearance Models Revisited, *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135-164, Also available online:http://www.ri.cmu.edu/pubs/pub_4601.html.

[13] T. Oggier, Michael Lehmann, M. S. Rolf Kaufmann, M. Richter, P. Metzler, G. Lang, F. Lustenberger and N. Blanc. An All-solid-state Optical Range Camera for 3d Real-time Imaging with Sub-centimeter Depth Resolution (swissranger), in: SPIE, *Conference on Optical System Design*, St. Etienne, September 2003.

[14] T. Oggier, B. Büttgen, F. Lustenberger, G. Becker, B. Rüegg and A. Hodac. Swissranger Sr3000 and First Experiences Based on Miniaturized 3d-tof Cameras, in: Proceedings of *1st Range Imaging Research Day*, September 8/9, 2005 at ETH Zurich, Switzerland, pp. 97-108, 2005.

[15] J. Picone. Speech Recognition: An Overview of Statistical Modeling of Acoustics, *JHU Summer School on Human Language Technology*, June 27, 2005.

[16] T. Wehrle and S. Kaiser. Emotion and Facial Expression, in Affective Interactions: Towards A New Generation of Computer Interface, ed. A. M. Paiva. Berlin: Springer-Verlag, pp. 49–63, 2000.



Christian Eckes received the diploma degree in Physics from the University of Dortmund in 1995 with a topic combining top-down and bottom-up segmentation cues in a interacting Potts spin system. From 1996 to 2000 he worked as a research assistant at the Institute for Neural Computation at the Ruhr-University Bochum and, as a visiting scholar, at the University of Southern California (USC), Los Angeles, U.S.A., respectively. In 2001 he joined the Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) in the NetMedia group as a research fellow and project manager working in the area of multimedia analysis. His research interests include human-computer interaction, multimedia analysis, pattern recognition and biological inspired computer vision.

Konstantin Biatov received master degree in applied mathematics from Moscow Institute Physics and Technology, Moscow, Russia in 1973 and Ph.D degree from Glushkov Institute of Cybernetics NAS Ukraine, Kiev in 1986. He was worked in the area of neural network, speech recognition and speech understanding in the Institute of Cybernetic from 1973 to 1998. From 1998 to 1999 he has been with Computer Research Laboratory of New Mexico State University in NM, USA where he was working as computer specialist III in lesser studied languages acquisition project and from 1999 to 2000 he has been with AT&T Bell Labs Research, Florham Park, NJ, USA where he was working as mathematical consultant in DARPA Communicator project. From 2001 he is in Fraunhofer IAIS in NetMedia group working in the area of audio data analysis for multimedia applications.

Frank Hülsken is a mathematician with background in pattern recognition. He is working as a research fellow at the Fraunhofer Institute for Intelligent Analysis and Information Systems. His research interests include speech recognition, image processing, 3D reconstruction and animation.

Joachim Köhler received his diploma and Dr.-Ing. degree in Communication Engineering from the RWTH Aachen and Munich University of Technology in 1992 and 2000, respectively. In 1993 he worked in the Realization Group of ICSI in Berkeley on robust speech processing algorithms. From 1994 until 1999 he worked in the speech group of the research and development centre of the SIEMENS AG in Munich. The topic of his PHD thesis is multilingual speech recognition and acoustic phone modelling. Since June 1999 he is with Fraunhofer IMK/IAIS in Sankt Augustin and head of the research group NetMedia. His personal research interests include speech recognition, spoken document and multimedia retrieval and MPEG-7 technologies.

Pia Breuer studied Computational Visualistics (Computervisualistik) at the University of Koblenz -Landau and completed her degree in 2006 with a diploma thesis about "Entwicklung einer prototypischen Gestenerkennung unter Verwendung einer IR-Tiefenkamera" at the Fraunhofer Institute for Media Communication IMK division NetMedia in Sankt Augustin.

Currently she works as a research assistant at the research group Media Systems at the University of Siegen. Her research interests include human-computer interaction and computer vision (automatic face reconstruction).

Pedro Branco graduated in Computer Science from University of Porto in 1997 and received the doctorate degree in Information Systems from University of Minho, in 2006. In 2000, he joined Fraunhofer's U.S. operations as Researcher/3D Software Engineer in the development of virtual reality interaction techniques. Since 2003 he has worked at IMEDIA in Providence, RI, studying user interface usability based on physiological monitoring. In January 2007 he joined the Department of Information Systems at University of Minho, Portugal as auxiliary professor. The topic of his doctoral dissertation is: "Computer-based Facial Expression Analysis for Assessing User Experience". His research interests are on monitoring users' facial expressions and the associated computer vision topics, intelligent user interfaces and anthropomorphic interfaces.

L. Miguel Encarnação is the Chief Technology Officer and Executive Vice President of the IMEDIA Inc., a Rhode Island, U.S. private research lab and technology transfer corporation focusing on the development and commercialization of innovative information and communication technologies in the area of interactive digital media. Dr. Encarnação is an internationally renowned expert on user-centered visualization design and human-computer interaction. Since 1992, he has been involved in design, development, performance analysis, test and evaluation, and program/project management of interactive graphics applications and data visualization systems.

Dr. Encarnação holds PhD, MS, and BS in Computer Science, and is currently serving as adjunct professor of Computer Science at the University of Rhode Island. Dr. Encarnação is the author or co-author of numerous contributions to peer-reviewed journals and conferences including Computers & Graphics, Presence and IEEE CG&A, and has made contributions to books on Computer Graphics education and programming. He is a member of the editorial boards of IEEE Computer Graphics & Applications, the International Journal of Technology and Human Interaction, and the International Journal for Virtual Reality, and a frequent reviewer for many technical journals as well as the U.S. National Science Foundation.