Gestural Interaction Using Feature Classification

Cornelius Malerczyk

ZGDV Computer Graphics Center
Rundeturmstrasse 10
64283, Darmstadt, Germany
cmalerc@zgdv.de

Abstract. This paper describes our ongoing research work on deviceless interaction using hand gesture recognition with a calibrated stereo system. Video-based interaction is one of the most intuitive kinds of Human-Computer-Interaction with Virtual-Reality applications due to the fact that users are not wired to a computer. If interaction with three-dimensional environments is considered, pointing, grabbing and releasing are the most intuitive gestures used by humans. This paper describes our video-based gesture recognition system that observes the user in front of a large displaying screen, identifying three different hand gestures in real time using 2D feature classification and determines 3D information like the 3D position of the user's hand or the pointing direction if performed. Different scenario applications like a virtual chess game against the computer or an industrial scenario have been developed and tested. To estimate the possible count of distinguishable gestures a sign language recognition application has been developed and tested using a single uncalibrated camera only.

1 Introduction

Hand gesture recognition in computer vision is an extensive area of research that encompasses anything from static pose estimation of the human hand to dynamic movements such as the recognition of sign languages. This paper describes our ongoing research work on deviceless interaction using hand gesture recognition with a calibrated stereo system. Video-based interaction is one of the most intuitive kinds of Human-Computer-Interaction with Virtual-Reality applications due to the fact that users are not wired to a computer. People frequently use gestures to communicate. Gestures are used for everything from pointing at an object or at a person to get their attention or to conveying information about space and temporal characteristics. Gesture recognition used for human computer interaction (HCI) comprehends both advantages and disadvantages depending on the technical constraints of the recognition system. One of the most important advantages is that no menu structure for an application is needed. Gesture recognition systems can be very powerful, for example in combination with speech recognition and speech synthesis systems. This multi-modal type of interaction with a computer is one of the most intuitive way for a human to communicate with a technical system. But there are also disadvantages using gesture recognition systems: Calibration inaccuracies of data gloves for example can destroy the performance of a system, this is often caused by a hand size problem. Further on, the user often needs to learn gestures,

while there is no self-explanatory interaction. Users are easily cognitively overloaded. Last but not least device driven recognition systems need a direct cable connection to the computer, which decreases the freedom of action of the user and leads to an uncomfortable feeling of the user. Therefore, we propose an easy and as far as possible self explanatory interaction with virtual environments using the three different static gestures: pointing, grabbing (closed hand) and releasing (opened hand) that are intuitive enough to be used by all kinds of different and even technically unversed users (see figure 1). The demands on the gesture recognition and tracking system proposed

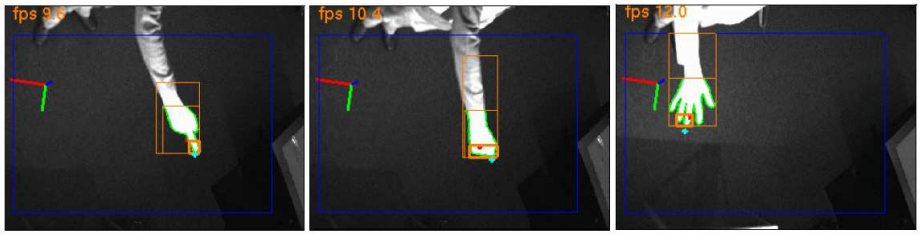

Fig. 1. Three different gestures seen by one of the cameras of the stereo system: Pointing (top), grab/closed hand (middle) and release/opened hand (bottom).

in this paper arise from the scenarios of interacting with virtual 3D environments itself. It is necessary to have a tracking system at hand that is able to handle the interaction of different users, no matter if they are left- or right-handed, if they use just the index finger for pointing or even the opened hand. In order to avoid new users are loosing patience, the training phase that has to be performed before the application starts has to be as simple and short as possible.

2 Related Work

Hand gesture recognition in computer vision is an extensive area of research that encompasses anything from static pose estimation of the human hand to dynamic movements such as the recognition of sign languages. A primary goal of gesture recognition research for Human-Computer-Interaction is to create a system, which can identify specific human hand gestures and use them to convey information or for device control. Therefore, hand gesture recognition systems can be divided into different tasks: Hand tracking, dynamic gesture recognition, static gesture recognition, sign language recognition and pointing. Elaborate reviews on vision-based hand pose estimation can be found in [Ero07] and [Coh07]. Due to the high amount of different approaches and methods for hand pose estimation, we only consider approaches, that are capable of tracking the human hand in real-time and differentiating static hand postures. Exemplarily, [Sch07] presented an approach for tracking the position and the orientation of four different postures of two hands in real-time by calculating the visual hulls of the hands on the GPU directly. Therefore, the results necessarily depend on the correct reconstruction of the 3D pose of the hand using at least three cameras. [Oha00] uses a

calibrated stereo system with two colour cameras to estimate the position and orientation of different hand postures in 3D. The approach is based on 2D feature extraction using various techniques based on geometric properties and template matching. Therefore, it is necessary to identify corresponding feature in image pairs and to derive 3D features that are afterwards fitted to an underlying 3D hand model to classify different gestures. Due to the fact that no orientation of the hand is determined in our approach, we are able to reduce the classification problem to pure 2D feature extraction and to calculate the 3D position of the hand for interaction purposes using the center of gravity of the segmented hand only.

3 System Calibration

The equipment for the gesture recognition system consists of one single standard PC, which is used for both rendering of the scenario applications and gesture recognition and tracking. A standard video beamer or a large plasma display is connected to the PC, displaying the application scenario in front of the interacting user. Two Firewire (IEEE1394) cameras are connected to the computer feeding the system with grey-scaled images of the interaction volume in real time. Lenses with additional infrared light diodes (without infrared light filters) are used to ensure a bright reflection of the human skin (see figure 2), which is necessary for the segmentation task described in section 4.
The purpose of the tracking system is to recognize and to track three different static gestures of the user (pointing, opened hand and closed hand) to enable intuitive interaction with the scenario applications. The approach is based on the recognition of the position of the human hand in 3D space within a self-calibrated stereo system [Aza95]. Therefore, position and orientation of the cameras are determined with respect to each other by swaying a small torch light for a few seconds in the designated interaction volume [SM02]. Afterwards only the world coordinate system has to be defined by marking the origin and the end of two axes of the world coordinate system. Within this coordinate system the corners of the displaying screen have to be declared to ensure a correct interpretation of especially the pointing gesture and its pointing direction. This calibration procedure has to be performed only once after setting up the cameras. During runtime of the system, difference images are used to detect moving objects, which then are analyzed and the 2D probabilities of a posture and its relevant parameters in 3D space is calculated. Smoothing of the tracking results like e.g. the 3D position of the hand using smoothing splines is used to reduce jittering effects [SE00], which leads to an immersing experience during the interaction without the need of any technical device.

4 Gesture Extraction

The segmentation of the user's hand is performed on 2D image basis. At the startup of the tracking system reference images of the empty interaction volume are taken from both cameras, smoothed using a 3*3 Gaussian filter mask and afterwards edge images are calculated using a standard 3*3 Sobel kernel. During runtime of the tracking system images captured by the stereo system are again smoothed and edge images are

calculated. These edge images are then compared with the edge reference images by calculating difference images. The resulting pair of images is afterwards segmented at a predefined threshold. Due to the camera setup at the left and right hand side of the user and due to the fact that the user is always interacting with a screen in front of him/her, it can be easily assumed that the lowest extracted segment larger then an adequate segment size can be chosen as the user's hand. Nevertheless, often not only the user's hand but also his/her forearm is extracted into one segment. Therefore, an approximately square rectangle at the bottom of the segment is chosen containing the final hand segment (see figure 2). This segment is used afterwards for feature extraction. For example the centers of gravity in both images are projected into 3D space to define the position of the user's hand in the world coordinate system. Accordingly, the lowest points of the segments are identified to be the finger tip if a pointing gesture is recognized.

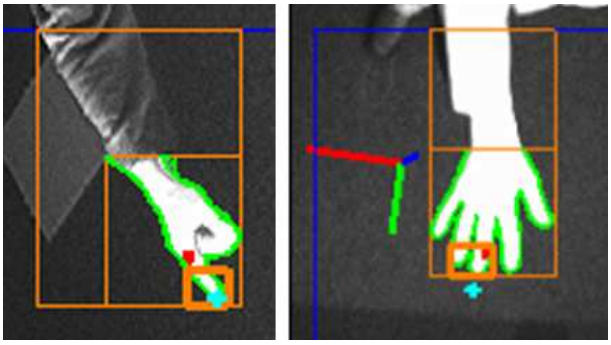

Fig. 2. Segmentation: Moving arm, hand extraction and index finger (left) and open hand (right) indicated by superimposed boxes.

5 Gesture Classification

For each frame pair of the stereo camera system several two-dimensional features of the segmented human hand (as described in the previous section) are extracted using standard image processing methods. Important examples of these parameters are:

- Boundary length of the segmented hand
- Ratio between boundary length and area of the segment
- Eccentricity of the segment
- Elongatedness of the segment
- Direction of the segment, if elongated
- Compactness of the segment
- Curvature of the segment boundary

Using two-dimensional feature extraction only has the advantage of a robust and fast parameter determination. However, parameters of one single object may differ in both

camera images. Therefore, all parameter pairs are sorted by their size and stored as a feature vector for classification of the gestures. Feature classification is a well known task for image understanding and object recognition. Often used classification methods are Hidden Markov Models (HMM) as e.g. described in [Sta96] and Artificial Neural Networks (ANN) as e.g. described in [Kje97] for the visual interpretation of hand gestures. For our system algorithms for machine learning for data mining tasks described in [Wit05] like the Naïve Bayes Classifier are used for gesture identification as e.g. described in [Gun05] for the analysis of expressive face and upper-body gestures. We tested the following four algorithms with respect to time of model calculation, classification rate and feature separation:

- Naïve Bayes Classifier
- Bayesian Network with K2-Hill-Climbing
- Sequential Minimal Optimization and
- Random Tree Classifier

A validation of the classification results of up to 800 feature vectors were performed using a standard cross-validation method using one half of the captured feature vectors for model construction and the other half to examine the recognition rate of the system. Our tests with different users showed that the Naïve Bayes Classifier leads to a sufficient balance for the model building process time of less then 5 seconds and an online classification rate of more than 50 Hz.

6 Gesture Training

To ensure a robust and stable recognition of the different gestures (with a classification result of more than 95%) a short training procedure for each new user of the system is necessary. Basically, this training procedure can be skipped using a large predefined training data set consisting of the training data of several users, whereas the recognition result decreases by up to 10% and therefore sporadic misinterpretations of the system may occur. For the training procedure the user is asked to perform the different gestures for approximately ten seconds each at the startup of the system. This procedure is easy and short enough so that the user does not lose interest in starting to interact with the application itself. If desired the training result may be individually saved for a later personalized usage of the tracking system.

7 Pointing Gesture

Using the hand-as-a-tool metaphor, hand gestures are used for mode changes. One should make a division between postures (static movement) or gestures (posture with change of position and/or orientation hand, or moving fingers). The term gesture is often used for both postures and dynamic gestures. Gesture recognition systems used for human-computer interaction require thorough feedback to reduce user's cognitive load. Interaction always needs a visual feedback directly and without delay on the output device. Since the human pointing posture is naturally not as precise as a technical device

like e.g. a laser pointer, it is important to permanently provide the user of the system with a visual feedback for comprehensible perception of his/her interaction instead of calculating the pointing direction as precise as possible. In opposite to both other used gestures the pointing posture demands on an extra parameter in 3D space: The pointing direction and therefore the intersection point of a pointing ray with the displaying screen is calculated by the definition of a target point. Due to the fact that the position of

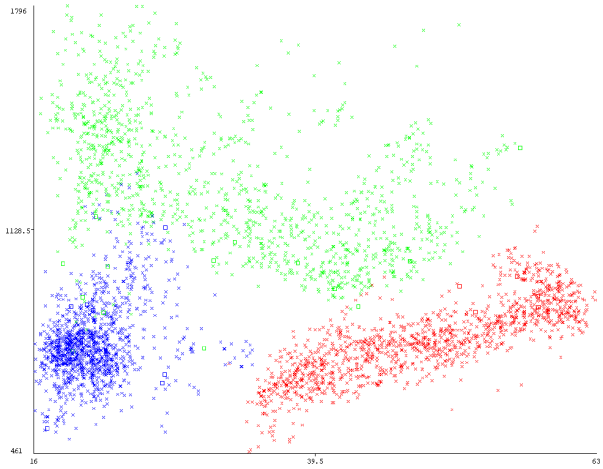

Fig. 3. Scatterplot of two selected features used for classification of pointing (blue), grabbing (red) and releasing (green).

the user is predefined (indicated by footstep markers on the floor), a 3D target point can easily be defined behind the user. The intersection of a ray starting at the target point and the direction defined by the currently reconstructed 3D position of the fingertip with the displaying screen is used to determine the position on the screen the user is pointing at.

8 Results

Using the hardware setup described in section 2 the system provides interaction feedback in real-time. After the calibration procedure performed only once at the system setup that needs approximately three minutes and after an initial training phase of three minutes (to perform the three different postures for ten seconds each and to build the classification model), the runtime recognition rates are at twenty frames per second and higher. The cameras deliver up to 30 frames per second with a delay shorter then 1/5 of a second. Even all image processing and computer vision tasks like segmentation, edge calculation and 3D reconstruction of corresponding image points last less then 50ms. Overall a recognition rate of up to 25 Hz is achieved. Nevertheless, as important is the

classification rate of the system. Figure 3 shows the classification results of one single user for three different gestures (pointing, grab/closed hand and release/open hand). The user performed each gesture 10 seconds during the training procedure, which leads to approximately 250 feature vectors for each gesture. Tables 1 and 2 show the recog-

Table 1. Classification matrix for three different gestures using a Naïve Bayes probabilistic model

Gesture performed vs. classified			
	Pointing	**Grab**	**Release**
Pointing	449	6	33
Grab	3	426	1
Release	24	2	379

nition and classification results for 60 seconds of interaction with a recognition rate of approximately 95% for new single feature vectors. Due to the fact that outliers (single incorrect classified gestures) can be eliminated using a post-processing queue a completely correct visualization of the performed gestures is achieved for interaction with the application scenario.

Table 2. Recognition rates for three different gestures

Training data	**Test data**	**Recognition rates**
Training session	Cross validation	95.1%
Testing session	Cross validation	97.6%
Training session	Testing session	94.7%

9 Applications

As a proof of concept two different scenario applications using virtual 3D environments for interaction have been developed. For a Virtual Chess application the user is standing in front of a large scaled screen rendering a three-dimensional chess board (see figure 4). Using the 3D position of the recognized gestures for grabbing and releasing the user is able to move chess pieces and therefore to play a game of chess against the computer. In the second scenario application the user is asked to place color labeled filters to virtual 3D industrial air pump system. An additional object snapping method is used to ensure a precise assembly of the three-dimensional objects, even if the user releases a filter object only roughly at the outlet of the air pump (see figure 4).

10 Sign Language Recognition

To estimate how may different gesture can be recognized and separated using the proposed methods a further application for the recognition and understanding of the American Manual Alphabet (AMA, American Sign Language Alphabet) has been developed.

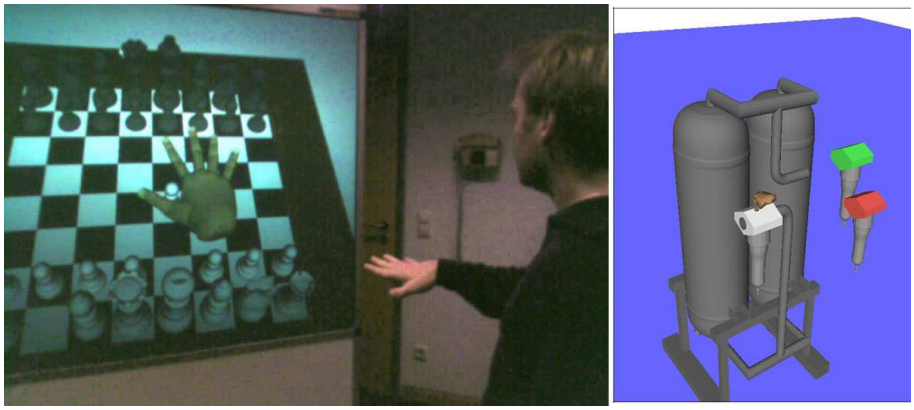

Fig. 4. Playing virtual chess against the computer (left) and Learning assistance scenario: Placing filters to an industrial air pump (right).

AMA is a manual alphabet that augments the vocabulary of American Sign Language (ASL) when spelling individual letters of a word is the preferred or only option. This application task is an obviously more complex scenario due to the fact that overall 26 different alphabetic characters have to be recognized and classified by the system. Due to the fact that all letters should be signed with the palm facing the viewer the usage of a single camera system like a standard webcam is sufficient. Without the special needs of a camera calibration procedure within a stereo setup as described in the scenarios above the most simple way to extract the hand shape is to use a simple color based webcam and a standard skin color segmentation of the captured image. In a first pre-processing step

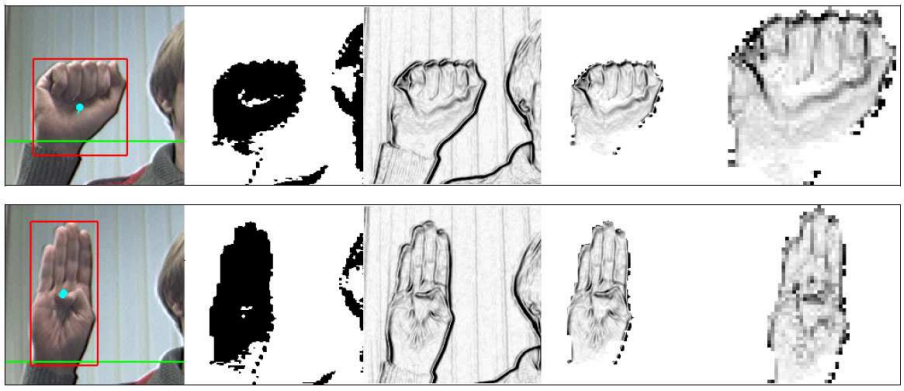

Fig. 5. Image processing pipeline for letters A (top row) and B (bottom): From left to right: Original camera image with augmentations, skin color segmentation, Sobel edge image, hand shape extraction, normalization of hand shape to 50*50 pixels.

the captured image is smoothed using a 3*3 Gaussian kernel. Afterwards, the image is converted from RGB color space to YCbCr[1] color space and a skin color segmentation within the intervalsand is performed. Assuming that the largest extracted segment seen in the image is the letter performing hand shape, the Sobel edge image of this segment is calculated. Finally, the extracted image is normalized to a given size of 50*50 pixels (see figure 5). For the classification procedure the normalized edge image of the hand shape (see figure 5, right) is used as the feature vector. Therefore, feature vectors for each frame are 2500-dimensional (50*50 luminance values). For classification a Bayesian Network with K2-Hill-Climbing is used, which leads to a model generation time of approximately one minute and an classification rate of up to 15 Hz. Due to the fact that the American Sign Language Alphabet uses the motion of the performing hand for the letters 'J' and 'Z' the center of gravity of the segmented hand shape within the captured image is analyzed in parallel (see figure 6) and used as an additional feature for classification. The system operates with up to 10 frames per second and with a recog-

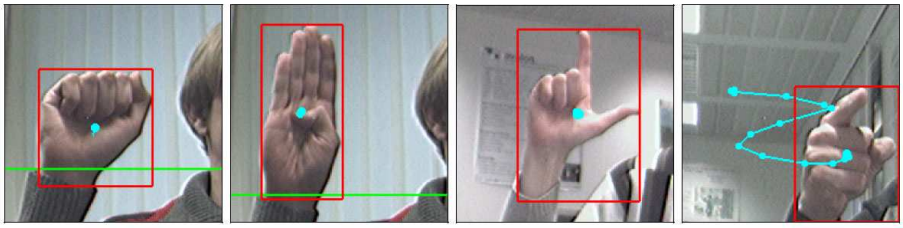

Fig. 6. ASL letters A, B, L, Z (from left to right), blue motion paths superimposed.

nition rate of 90% ($\pm$4%). Incorrect classifications appear mainly for visually similar letters like 'M' and 'N' (32/37 errors) or 'V' and 'K' (48/39 errors). To compensate incorrect recognized letters during finger spelling a "predicted text"-algorithm is used, which is basically well known from entering text (like SMS, Short Message Service) on the keypad of a mobile phone.

11 Conclusion

In this paper we presented a video-based gesture recognition system using a calibrated stereo system with two off-the-shelf cameras, which is able to identify three different hand gestures (pointing, grabbing and releasing) and determine the relevant parameters like 3D position of the hand and the pointing direction to enable an easy to learn and intuitive interaction with 3D virtual environments. Only a short training phase is needed to ensure high recognition rated of more than 95% in real-time. Two different scenario applications have been developed and tested addressing the application domains of entertainment and learning assistance. Furthermore, an application for sign

[1] Defined in the ITU-R BT.601 (formerly CCIR 601) standard for use with digital component video

language recognition has been developed to estimate the limit of distinguishable gestures using the proposed classification methods. Using one uncalibrated camera only, the recognition rate is up to 90% at a frame rate of up to 10 frames per second.

12 Acknowledgments

Parts of the work presented here were accomplished with support of the European Commission through the SIMILAR (www.similar.cc) Network of Excellence (FP6-507609).

References

Azarbayejani, A., Pentland, A.: Camera self-calibration from one point correspondence. Media Lab Technical Report 341, 1995.

Cohen, C., The Gesture Recognition Home Page - A brief overview of gesture recognition, Website, retrieved Oct 2007, http://www.cybernet.com/ ccohen/

Erol, A. et. al., Vision-based hand pose estimation: A review, Computer Vision and Image Understanding, Volume 108, Issues 1-2, October-November 2007, pages 52-73,Special Issue on Vision for Human-Computer Interaction

Gavrila, D., The Visual Analysis of Human Movement: A Survey. Computer Vision and Image Understanding, 73(1): 82–98, Jan. 1999.

Gunes, H., Piccardi, M.: Bi-modal affect recognition from expressive face and upper-body gesture by single frame analysis and multi-frame post integration, National ICT Australia HCSNet Multimodal User Interaction Workshop, Sydney, 1314 September 2005.

Kjeldsen, F.: Visual Interpretation of Hand Gestures as a Practical Interface Modality, Dissertation, Columbia University in the City of New York, 1997.

Kohler, M.: Vision based hand gesture recognition systems. University of Dortmund,Website, Retrieved October 2007 from http://ls7-www.cs.uni- dortmund.de/research/gesture/.

Malerczyk, C.: Interactive Museum Exhibit Using Pointing Gesture Recognition. In: Skala, Vaclav (Ed.) u.a.; European Association for Computer Graphics (Eurographics):WSCG 2004. Short Communications Volume II. Plzen : University of West Bohemia, 2004, S. 165-171

Malerczyk, C., Daehne, P., Schnaider, M., Exploring Digitized Artworks by Pointing Posture Recognition, 6th International Symposium on Virtual Reality, Archaeology and Cultural Heritage, Pisa, Italy, 8th - 11th November 2005.

O'Hagen, R., Zelinsky A., Visual Gesture Interfaces to Virtual Environments, Proceedings if AUIC2000, Canberra, Australia, 2000

Sun, S., Egerstedt M.: Control theoretic smoothing splines. IEEE Transactions on automatic control 45, 12 (2000).

Schlattman, M., Klein, R., Simultaneous 4 gestures 6 DOF real-time two-hand tracking without any markers, Proceedings of the 2007 ACM symposium on Virtual reality software and technology, pp 39-42, Newport Beach, California, USA, 2007

Schwald, B., Malerczyk, C.: Controlling virtual worlds using interaction spheres. In Proceedings of 5th Symposium on Virtual Reality (SVR) 2002, C.A. Vidal B. C. S., (Ed.), pp. 3?14.

Starner, T., Pentland, A.: Real-Time American Sign Language Recognition from Video Using Hidden Markov Models, Technical Report 375, Massachusetts Institute of Technology Media Laboratory, Cambridge, 1996.

Witten, I., Frank, E.: Data Mining Practical machine learning tools and techniques, Second Edition, Morgan Kaufmann, San Francisco, 2005.