

Dynamic Gestures as an Input Device for Directing a Mobile Platform

M. Ehrenmann, T. Lütticke and R. Dillmann
University of Karlsruhe
Institute for Process Control & Robotics
Karlsruhe, D 76128, Germany

Abstract

Giving advice to a mobile robot still requires classical user interfaces. A more intuitive way of commanding can be provided by verbal or gesture commands. In this article, we present new approaches and enhancements for established methods that are in use in our laboratory. Our aim is to direct a robot with simple dynamic gestures. Within this paper we will concentrate on visual gesture recognition.

Based on skin color segmentation algorithms for tracking the user's hand, hidden Markov models will be used for gesture type recognition. Filters applied to the recorded trajectory strongly compress the input data. They also mark start and end point of a possible gesture. The hidden Markov models have been enhanced by a threshold model in order to wipe out insignificant movements. Pre-classification of the reference gestures serves for keeping computational effort low.

1 Introduction

Nowadays, mobile robots are being handled mostly via graphical user interfaces in standard PCs, PDAs or teachpanels. Aiming to facilitate man machine interaction in a direct and intuitive way, auditive and gesture based interfaces have been a very important research topic in recent years. According to movements performed by humans when instructing car drivers to park, usage of dynamic gestures for the instruction of mobile platforms is being investigated at the Institute for Real-Time Systems and Robotics. The only significant input for the interpretation of user actions should be the trajectory of one of the user's hands. Finger poses or wrist angles shall not

influence the classifications.

Main problems in this variation of gesture recognition can be summarized as follows:

- Fast and robust tracking of user actions for extracting the hand trajectory.
- Training of model parameters with several users.
- Classification of the recorded trajectories with respect to trained model gestures.

A substantial aspect of the last matter is to prevent the mapping of insignificant movements to a reference model. Furthermore, classification should be accomplished in real-time.

1.1 State of the Art

We will briefly discuss present methods addressing these kinds of difficulties. Image based tracking of the human hand is considered as a special and very demanding problem. The many degrees of freedom and the complex kinematic structure entail not only the possibility of shape and brightness changes but also of occlusions. This complicates exact mapping to particular fingers. Thus, image based classification of gestures often does not rely on hand postures.

Hand Tracking: Approaches for human hand tracking are based on silhouette projections [15, 4], information about movement [21] or colored markers. Some applications aim at entire 3D reconstruction with complex hand models, but doing this in real-time is a hard task [17]. In most cases, no information about finger postures is necessary. Computation rather relies on optical flow [11], correlation of stereo color images [1] or skin color segmentation [18].

Gesture Recognition: “Gesture” is the commonly used term for a symbol used in order to command, instruct or converse what is expressed through hand postures or hand movements. Preferred sensors for the detection of gestures are image based systems or data gloves. An overview of several gesture recognition systems has been compiled by Kohler [8]. Here, only vision based systems are being taken into account. Hand signs can be divided into static and dynamic gestures, depending on whether the significant part is either a finger posture or hand movement.

Static Gestures: Kestler [6] uses camera images in order to recognize and classify static hand gestures. After preprocessing the images, tensors of gesture patterns are compared with the transformed actual images similar to an Eigenspace method. When applying contour models for hand tracking, parameters of the current feature distribution can be used directly to describe a gesture [2, 5]. In essence, this approach is similar to matching elastic graphs to image features proposed in [20].

Dynamic Gestures: A system for recognition of korean sign language has been realized using data gloves, magnetic field based position sensors with fuzzy min-max neural networks [7]. Here, finger posture as well as hand movements are significant for gesture classification. However, it can be observed that nowadays hidden Markov models are employed more and more. One reason for this is that neural networks can hardly detect insignificant trajectories [19]. These models are mostly being trained and tested with standard methods (introductions can be found in [3, 14]). Very good results with hidden Markov models have been reached by Lee performing dynamic gestures to control a slide presentation [9]. In order to prevent false classifications, a threshold model is introduced. Since the problem of dynamic gesture recognition is similar to hand writing recognition, the same approaches can be used in either case [13]. Recently, cameras are being mounted on the performer in order to observe his hand movements [12].

2 Experimenting System

In the following, the image processing and experimenting system at our institute and our approach based on hidden Markov models will be discussed.

Before feeding trajectories into a classification system, observation of the user’s hand must be performed. This is to be realized visually without using any markers.

2.1 System components

The sensor employed for hand tracking is not yet installed on a mobile platform. So far, it consists of a turn-tilt unit shipped by Amtec. Mounted on top are two Sony 777AP color cameras (see fig. 1). Digitizers are two standard Matrox MeteorII framegrabbers.

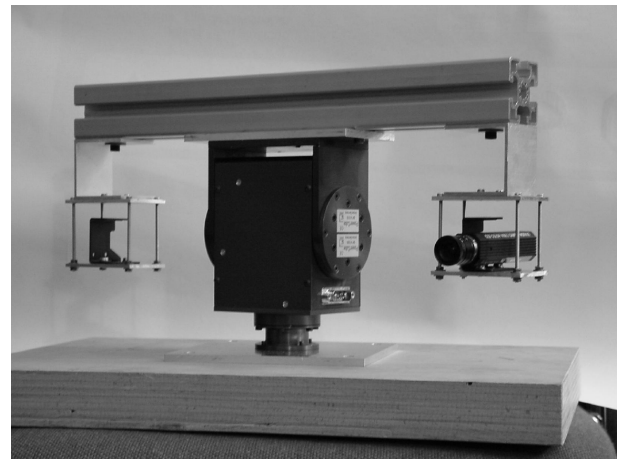


Figure 1: Camera head of Experimenting system.

2.2 Image acquisition

Detection and tracking of the user’s hand is based on skin color segmentation (see f.e. [22]). Incoming *RGB* images are being binarized after converting them into *HSI* color space. As thresholds under artificial light laboratory conditions, we use 3 and 31 for the lower and upper limit of the *H* value. After this, a closing filter is applied in order to fill gaps in the resulting binary image.

Before performing, only one person is allowed to stay in the camera’s scope. Then, three resulting skin color regions can easily be mapped to the user’s head and hands (see fig. 2). Among the latter ones, the left or right hand can be chosen for tracking.



Figure 2: Image Processing Steps (Grabbed image, Color Segmentation, Closing Operation).

The use of local windows and multithreading approaches speed this steps up to 15Hz on a double PentiumIII system with a 500MHz clock grabbing two full NTSC images.

2.3 Trajectory filters

The determined sequence of centers of gravity of the hand segment is fed into a multilevel filter for smoothing and data reduction purposes. More exact, the point sequences first serve as input for a neighbourhood filter. This filter rejects successive closely lying points and computes direction vectors between the pixel coordinates.

Yet there cannot be a two-dimensional input for hidden Markov models, direction vectors of the segments are mapped onto an 16 element alphabet (see fig. 3). On the one hand, this number is large enough to represent the performed movements. On the other hand, it is small enough to restrict computational effort.

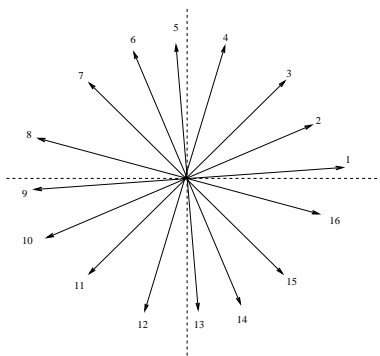


Figure 3: Codebook used for Vector Quantization with 16 words.

This sequence is reduced furthermore applying an

identity filter. Since every direction index results from a direction vector representing it's orientation, one index is sufficient for coding the vector's particular direction of movement. Several subsequent identical indices are redundant and can be mapped to a singular one.

Altogether, a reduction of the input sequence from 14% up to 96% can be reached depending on the gesture type. An example is depicted in figure 4.

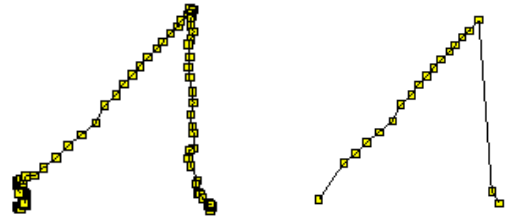


Figure 4: Recorded and filtered Trajectory.

A fundamental result for action recognition is that human beings perceive actions as a sequence of clearly separated single acts. The most relevant information for the interpretation of an action lies in the state of change inbetween two such single actions [10]. In order to separate start and end point of a gesture A start and stop recognizer is applied before the filters. It checks for short stops of the hand. This event serves as trigger detecting start and end point of the trajectory. A continually running interpretation as described in [9] was rejected because of higher computational effort. When interacting with the system, this kind of filter does not claim unintuitive postulations.

3 Gesture classification

Training of the hidden Markov models is done with the Baum-Welch algorithm which is a well established method. Hereby, we make use of left-right models with a jump delimiting $\Delta = 2$. For testing, five reference gestures have been selected, each trained with 10 distinct examples (see fig. 5).

With the techniques introduced in the above sections, the overall recognition system can be described. The entire processing is organized as depicted in figure 6.

A recorded trajectory can now be classified in four different manners:

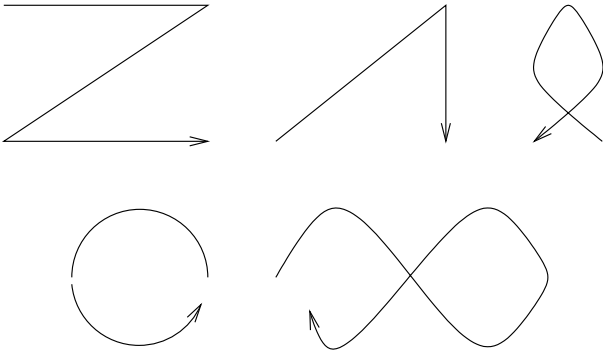


Figure 5: Reference gestures.

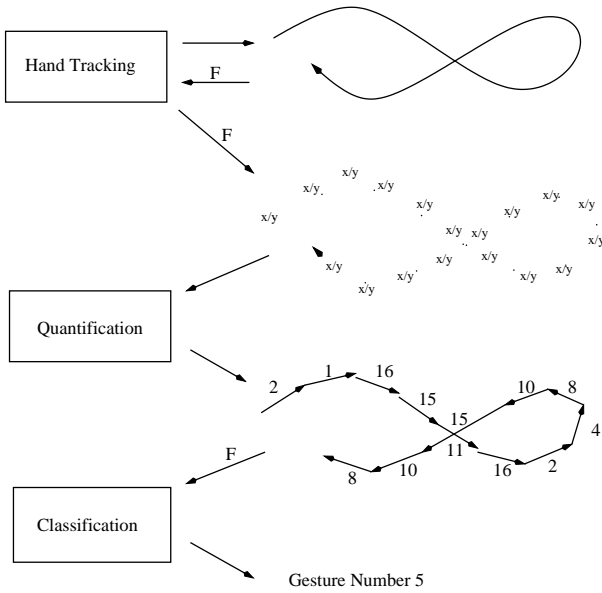


Figure 6: Gesture recognition process with its single phases. The employment of filters is marked by “F”.

Maximum probability (MP): Testing for maximum probability means feeding the recorded trajectory into each reference model and computing the according probability. The one with the highest output is being chosen. Computation of the maximum is being done with the Viterbi algorithm [16].

Threshold model (TM): Before classifying, a threshold model is being constructed fusing all reference models. The trajectory is fed into the reference models as well as into the threshold model. If the highest result emitted from one of the reference models exceeds the threshold model output, this reference model is accepted

as recognized gesture. If the reference model values are all below the threshold model value, the gesture is rejected being meaningless. This approach has first been proposed in [9].

Hierarchical classification with maximum probability (HP): The gestures get subdivided according to their length in several complexity classes. Classification with maximal probability does only take reference models of this class into account.

Hierarchical classification with threshold model (HT): Here, a class subdivision is being applied as well. For every class, a threshold model is computed, which determines acceptance or rejection.

New to this approach is the supplementation of the threshold model with several classes lowering computational effort. The classification for the proposed set of reference gestures is shown in figure 7.

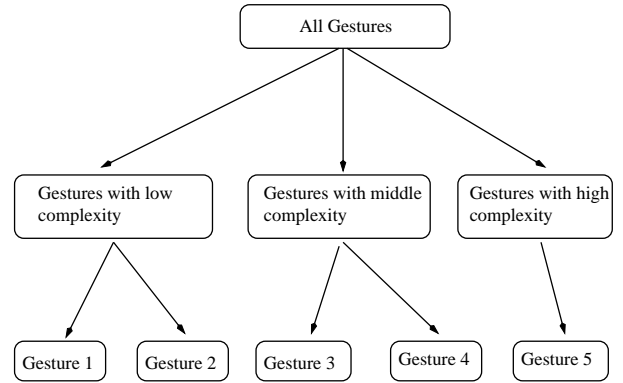


Figure 7: Classification of reference gestures according to their complexity.

4 Experimental results

The tests being run with more than 200 demonstrations have shown the results summarized in table 1.

We have to remark that classification with a simple selection of the reference model emitting the highest value (MP) shows best results but is useless for everyday use because of its incapacity to reject meaningless trajectories. The threshold model is demanding enough to be able to guarantee not to accept an insignificant action as a gesture. Since the effort of classification with the hierarchical approach (HT) is less

Geste	MP	TM	HP	HT
1	100%	77%	100%	68%
2	93%	79%	100%	83%
3	100%	76%	80%	84%
4	100%	85%	88%	73%
5	100%	100%	82%	73%
Gesamt	99%	83%	89%	76%

Table 1: Testing results with according classification variant.

than the third part of the threshold model approach we have decided to continue our work with this technique.

5 Conclusion and future enhancements

In this article approaches have been presented that realize the tracking of human hand movements and their classification as dynamic gestures. Classification itself is done on basis of hidden Markov models. In order to reject insignificant movements, the threshold model has been implemented as a supplement. A new contribution to this kind of classification is the classification of reference gestures into several complexity classes. This way, computational effort can be lowered significantly. Another new technique is the simple and effective approach for triggering the recognition process. This is done by the described start/stop filter. Additional filters reduce input data for the recognizer dramatically thus speeding up classification furthermore.

Enhancements of our systems will address image processing in first place. The employed skin color segmentation is not working properly with natural light. Thus, adaptive methods will be used. A second point will be hand tracking in 3D. The observed trajectories can then be mapped on a hyperplane where classification is done. Here, saccade movements will augment the space for possible movements.

ACKNOWLEDGMENT

This work has partially been supported by the BMBF project “Morpha”. It has been performed at the Institute for Real-Time Computer Systems &

Robotics, Department of Computer Science, University of Karlsruhe.

References

- [1] A. Arsenio and J. Santos-Victor. Robust visual tracking by an active observer. In *Proceedings of the IEEE International Conference on Intelligent Robots and Systems (IROS)*, volume 3, pages 1342–1347, 1997.
- [2] A. Blake and M. Isard. *Active Contours*. Springer, 1998.
- [3] G. Forney. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, März 1973.
- [4] D. Gavrilu and L. Davis. Towards 3d model-based tracking and recognition of human movement: a multi-view approach. In *International Workshop on Face and Gesture Recognition, Zürich*, 1995.
- [5] T. Heap and F. Samaria. Real-time hand tracking and gesture recognition using smart snakes. Technical report, Olivetti Research Limited, 20. Juni 1995.
- [6] H. Kestler, M. Borst, and H. Neumann. Einfache Handgestikerkennung mit einem zweistufigen Nearest-Neighbour Klassifikator. Technical report, Universität Ulm, SFB 527, 96/6, 1996.
- [7] J. Kim, W. Jang, and Z. Bien. A dynamic gesture recognition system for the korean sign language (KSL). *IEEE Transactions on Systems, Man and Cybernetics-Part B: Cybernetics*, 26(2):354–359, April 1996.
- [8] M. Kohler. *Übersicht Handgesten-erkennung*. <http://ls7-www.cs.uni-dortmund.de/research/gesture/>, 2000.
- [9] H. Lee and J. Kim. An HMM-based threshold model approach for gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(10):961–973, Oktober 1999.
- [10] D. Newton and et al. The objective basis of behaviour units. *Journal of Personality and Social Psychology*, 35, 1977.
- [11] P. Nordlund and T. Uhlin. Closing the loop: Detection and pursuit of a moving object with sensors onboard a moving robot. *Image Vision and Computing*, 14(4):265–275, 1996.

- [12] A. Pentland. Looking at people: Sensing for ubiquitous and wearable computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):107–119, Januar 2000.
- [13] R. Plamondon and S. Srihari. On-line and off-line handwriting recognition: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):63–84, Januar 2000.
- [14] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–285, Februar 1989.
- [15] J. Rehg and T. Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *ECCV*, pages 35–46, 1994.
- [16] M. S. Ryan and G. R. Nudd. The viterbi algorithm. Warwick Research Report RR 238, Department of Computer Science, University of Warwick, Coventry, Februar 1993.
- [17] N. Shimada and Y. Shirai. 3d hand pose estimation and shape model refinement from a monocular image sequence. In *Proceedings of the VSMM, Gifu*, pages 423–428, 1996.
- [18] H. Sidenbladh, D. Kragic, and H. Christensen. A person following behaviour for a mobile robot. In *Proceedings of the IEEE International Conference on Robotics and Automation, Detroit, MI, USA*, pages 670–675, April 1999.
- [19] T. Starner. Real-time american sign language recognition from video using hidden Markov models. In *Proceedings of the IEEE International Symposium on Computer Vision*, pages 265–270, 1995.
- [20] J. Triesch and Chr. von der Malsburg. Robotic gesture recognition. In *Proceedings of the Bielefeld Gesture Workshop*. Springer, 17.-19. September 1997.
- [21] M. Yamamoto and K. Koshikawa. Human motion analysis based on a robot arm model. In *CVPR*, pages 664–665, 1991.
- [22] J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaptation. In *Proceedings of ACCV, Hong Kong*, volume 2, pages 687–694, 1998.