PROJECT	A GESTURE BASED TOOL FOR STERLILE BROWSING OF RADIOLOGY IMAGE
TEAM ID	PNT2022TMID39067

Data Gathering for Gesture Recognition Systems Based on Single Color-, Stereo Color- and Thermal Cameras

Abstract

In this paper, we present our results of automatic gesture recognition systems using different types of cameras in order to compare them in reference to their performances in segmentation. The acquired image segments provide the data for further analysis. The images of a single camera system are mostly used as input data in the research area of gesture recognition. In comparison to that, the analysis results of a stereo color camera and a thermal camera system are used to determine the advantages and disadvantages of these camera systems. On this basis, a real-time gesture recognition system is proposed to classify alphabets (A-Z) and numbers (0-9) with an average recognition rate of 98% using Hidden Markov Models (HMM).

Keywords: *Gesture Recognition, Stereo Camera System, Thermal Camera, Computer Vision & Image Processing, Pattern Recognition.*

1. Introduction

Gesture recognition is an important area for novel human computer interaction (HCI) systems and a lot of research has been focused on it. These systems differ in basic approaches depending on the area in which it is used. Basically, the field of gestures can be separated into dynamic gestures (e.g. writing letters or numbers) and static postures (e.g. sign language). The goal of gesture analysis and interpretation is to push the advanced human-machine communication in order to bring the performance of human-machine interaction closer to human-human interaction. The most important component of gesture recognition systems is the exact segmentation and recognition of the hands and the face which depends on the data gathering. Therefore, different types of cameras are established in this research area (e.g. single color-, stereo color-, thermal cameras).

Most researchers use single color cameras for data acquisition . A big advantage of these cameras is that they are fast and simple to control, so it is possible to realize a suitable gesture recognition system also in real-time applications. Additionally, the color map of the image can be used for skin color recognition in order to improve the segmentation results. However the robustness of such a system can suffer from a complicated real background. And the separation of region of interest will still be a challenging problem if only one camera system is used.

Binh et al. used a single color camera for data acquisition and afterwards pseudo two dimensional Hidden Markov Models (P2-DHMMs) to recognize posture with good results. They used Kalman filter and blobs analysis of hands for hand tracking. Under controlled conditions they achieved a recognition rate of 98% for the classification of 36 postures of ASL (American Sign Language) in real-time. However, a slow movement of gesture is necessary and occlusions between hand and face have to be avoided. Also, wearing long sleeves and the presence of a homogenous background are preconditions in their system which cannot always be assured in a real-time application. Liu et al. developed a gesture recognition system to recognize 26 alphabets from a database by using different HMM topologies. They achieve the best recognition rate of 89.6% by using the Left-Right topology. Stereo cameras are rarely used in the field of gesture recognition. Their advantages are often ignored or rather not utilized. Malassiotis et al. had demonstrated a complete stereo based system for the recognition of 20 static hand postures from the German sign language alphabet based on a 3D sensor. They obtained the best results using principle component analysis (PCA) approach based on 3D pose compensation. Also Elmezain et al. has been working with different HMM topologies for gesture recognition by using a stereo color camera system. They analyzed 36 isolated gestures and achieved the best results using LRB topology.

In the last few years, thermal cameras are often used in the field of face recognition , E.g. Socolinsky et al. proposed the combination of thermal and visual cameras for face recognition. Their analysis shows that thermal cameras achieved the similar performance as normal visual cameras. And thermal cameras have their own advantages under uncontrolled illumination conditions (indoor and outdoor scene). Still, theuse of thermal cameras for the segmentation in the field of gesture recognition is rare. Based on that, this paper will give a fundamental analysis of using different camera systems for the data gathering.

This paper introduces a novel system to recognize continuous hand gestures from alphabets A-Z and numbers from 0-9 in real-time by using the motion trajectory of a single hand with the aid of HMM. For segmentation, three different types of cameras (single color-, stereo color- and thermal camera) are analyzed for their advantages and disadvantages. These segmentation results build the basis for the feature extraction and the tracking of hands and face. The orientation between two following points was extracted and used as a basic feature for HMM. These HMM are trained by Baum-Welch (BW) algorithm and the sequences are analyzed by Viterbi path. We have built a database with at least 30 video sequences per gesture (A-Z & 0-9) and used 20 sequences for training and 10 sequences to test our algorithm. Additionally, we have accomplished a lot of online tests.

This paper is organized as follows: advantages and disadvantages of the three different camera types for segmentation are analyzed and evaluated in section 2. In section 3, the real-time gesture recognition system and experimental results are presented. Finally, the summary and conclusion are presented in section 4.

2. Gesture recognition system

The proposed system in this paper has three processing steps. The first step is the segmentation which depends on the evaluation of data gathered by different camera systems. Then, the image segments of the first step provide the data input into the feature extraction and tracking process. Then the third step of this system is the classification process using Hidden Markov Models.

2.1 Data gathering and evaluation

In the field of the segmentation, different types of cameras are used for data gathering (Fig. 1) to build up a gesture recognition system. The single color camera is the most common type of data acquisition tool because of its ease of use and fast data evaluation capability even with high resolution images. The stereo image evaluation forms another approach. In addition to the color information, the depth information, which is determined through a disparity calculation, is used. Nevertheless, the disadvantage of stereo calculation is the increased computational cost. The thermal forms the third type of data production. In this case, the temperature of an object is captured with the help of infrared radiations and afterwards shown in the image.

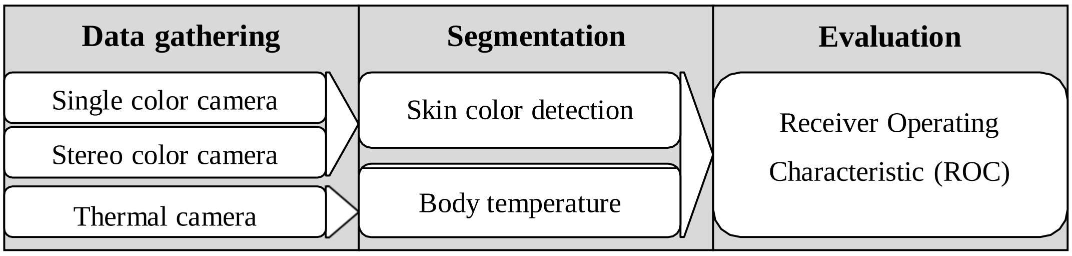


Figure 1. Data gathering and evaluation for gesture and posture recognition

2.2. Segmentation

Single/Stereo color camera. For skin color segmentation in colored images, at first two decisions must be made. The choice of the color space is the most important feature. The next allocation of a pixel as skin or non skin is only a problem of classification. Therefore a suitable skin color model is generated to check the affiliation to a skin color class. The most suitable color spaces are the ones which are oriented towards perception, because they are close to the human perception system in the way that the color- and brightness information are separate from each other.

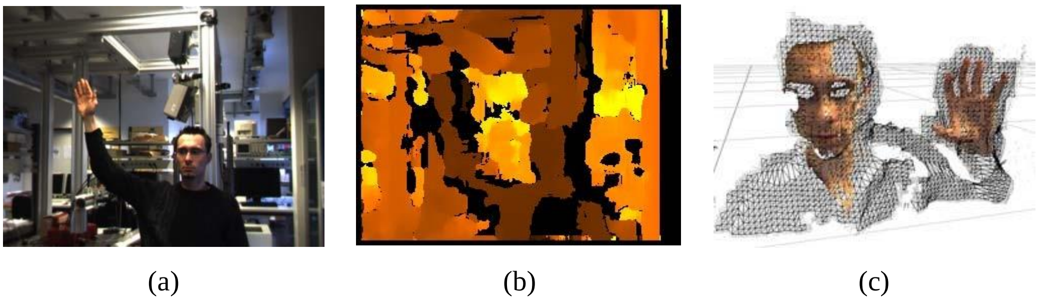
In our approach we use YC_bC_r color space where Y represents brightness and C_bC_r refers to chrominance. We ignore Y channel in order to reduce the effect of brightness variation and use only the chrominance channels which fully represent the color. The human skin color is found in a small area of the chrominance plane; so a pixel can be classified as skin or non skin by using a Gaussian model. A large database of skin pixels is used to train the Gaussian model, which is characterized by the mean vector μ (eq. 1) and covariance matrix σ (eq. 2). Since are skin color model is based on the chrominance plane C_b and C_r, a two dimensional vector with $x_i=[C_bC_r]^T$ was used as an input. The Mahalanobis-Distance (eq. 3) describes the distribution of an elliptical area with the highest probability with a mean value μ . The probability p (eq. 3) is continuously decreasing whereas orientation and slope will be described by the values of the covariance matrix σ.

$$\mu = \frac{1}{n-1}\sum_{i=1}^{n} x_i \qquad (1)$$

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)(x_i - \mu)^T \qquad (2)$$

$$p(y \mid j) = \frac{1}{2\pi \cdot \sqrt{|\sigma_j|}} \cdot \exp\left(-\frac{1}{2} \cdot (y - \mu_j)^T \sigma_j^{-1} (y - \mu_j) \right) \tag{3}$$

Now, all skin-colored areas can be segmented using the skin color model in the image. This already shows the first deficit in single cameras. With an inhomogeneous background, like in figure 2(a), it is not possible to segment the hands and the face perfectly. Further, it is not possible to separate ambiguously overlapping face and hands. These disadvantages can be overcome with a stereo camera system by using depth information. The camera calibration data and the image features that are matched based on the cross correlation of the left and the right images can be used to estimate the depth information of a 3D point P(X, Y, Z). Also an unequivocal separation of the user from an inhomogeneous background is possible by utilizing the depth map (Fig. 2b). Furthermore, the hands can be held in front of the face and all areas are assigned ambiguously to each other. However, this approach introduces some problems. Large areas like a closed hand or a head etc. can be easily segmented, but it can be problematic to get depth information of smaller areas like single spread fingers, which depends on the calculation of disparity.

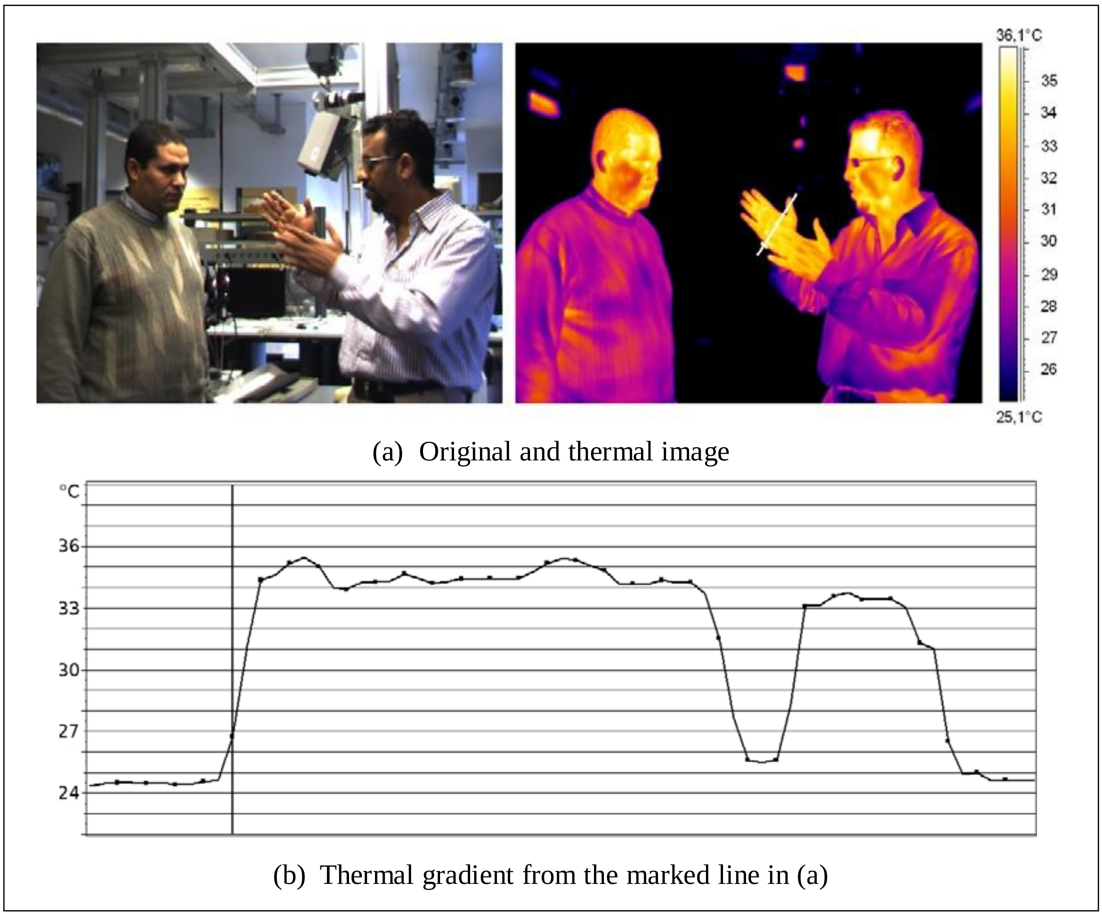


(a) (b) (c)

Figure 2. (a) Original single/stereo color camera (b) depth information (c) 3D structure from a stereo camera

Thermal camera. An infrared camera is a device that detects infrared radiation (temperature) from the target object and converts it into an electronic signal to generate a thermal picture on a monitor or to make temperature calculations on it. The temperature which is captured by an infrared camera can be measured or quantified exactly, so that not only the thermal behavior can be observed but also the relative magnitude of temperature-related problems can be recognized and noted.

In figure 3a, a normal scene of human interaction captured by a thermal camera is shown. Normally, the background can be neglected because the human temperature can be found in a small thermal area. As shown in figure 3a, the areas of the head and the articulating hands can be well separated from the background. Besides, the objects have very sharp contours and it even allows a good and clear segmentation of very small areas. The graph in figure 3b shows the temperature course of the straight line in figure 3a. Clearly, the sharp edges and the exact area of the hand are recognizable. However, like in the single camera case, the information of overlapping of hands or the face is still not possible to extract due to the missing depth information.



(a) Original and thermal image

(b) Thermal gradient from the marked line in (a)

Figure 3. Solution from thermal camera

2.3. Evaluation of segmentation results

For the proposed gesture recognition system, three different types of cameras are reviewed. In this manner, our experimental data is captured synchronously by a single color, a stereo color and a thermal camera system in complex real situations where the system had to recognize different types of gestures. The evaluation occurs by using the *Receiver Operating Characteristic (*ROC) curves. Therefore, the required ground truths of real skin pixels were marked by hand.

Figure 4a shows a frequency distribution which is separated into two classes (in this case the areas of 'skin' and 'non skin') by using a threshold. TP indicates the sum of *True Positive* pixels, which are coming from the ground truth and identified as a skin pixel by the system. Also FN is defined as the sum of *False Negative*, FP as *False Positive*, and TN as *True Negative* pixels. As in most cases no unequivocal differentiation of the classes is possible by using a normal threshold. If the threshold is getting smaller the number of false as positive (FP) values decreases, meanwhile the number of FN increases. The effect of shifting the threshold can be represented by the Receiver Operating Characteristic curves by using the *True Positive Rate* (TPR), the *False Positive Rate* (FPR) and the *Accuracy* (ACC), which can be calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{4}$$

$$FPR = \frac{FP}{FP + TN} \tag{5}$$

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} \tag{6}$$

In figure 4b an example of different types of classifiers is presented, whereas the curves differ significantly in the curvature. Furthermore, different working points A_1, *Error Equal Rate* (ERR) and A_2 are marked. A_1 is a working point where a high TPR with a low FPR exists. In contrast to A_1 the point EER describes the area where no value (TPR or FPR) is preferred, i.e. a theoretical optimum. This optimum is nearby the 45° line. In general, the more the working point gets to 100% TPR and 0% FPR the better the recognition is.

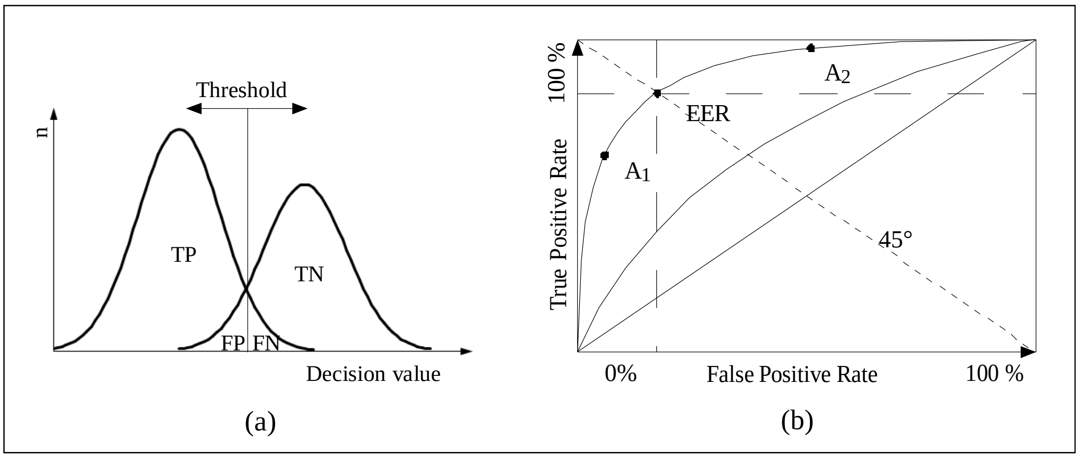
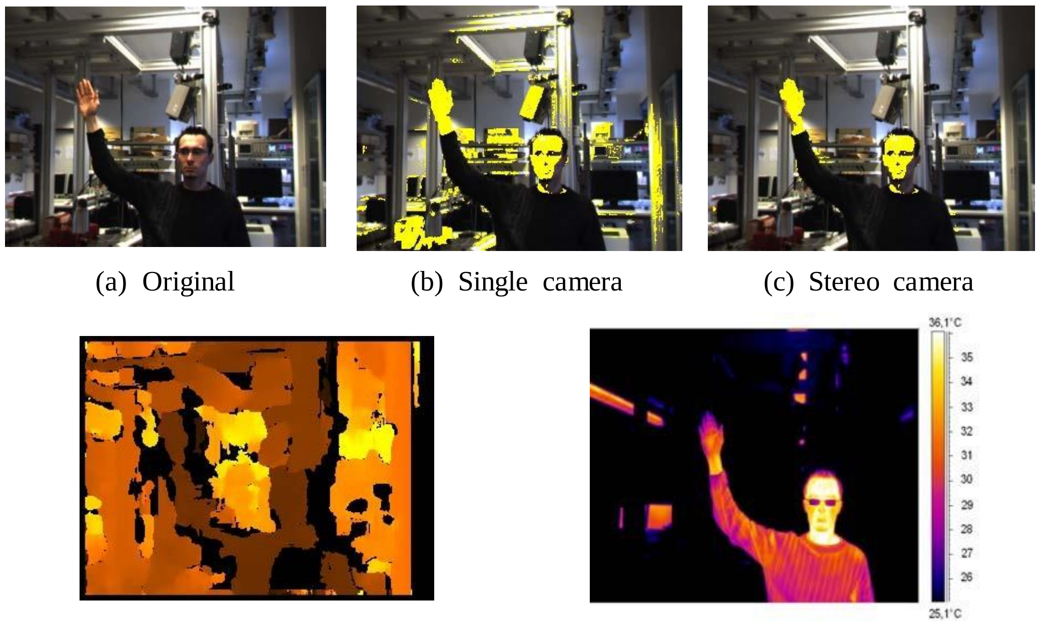


Figure 4. (a) shows a frequency distribution which is seperated into two classes and (b) shows different Receiver Operation Characteristic (ROC) curves

Figure 5a illustrates an image of our database captured by a single camera. Thereby, parts of the background are also segmented by the skin color model (Fig. 5b). Hence, in comparison to the other camera types under consideration, the single camera system has a relatively low mean TPR of 79.86%, ACC rate of 92.56% and a high mean FPR of 7.07%. Without any priori or additional information this is not enough for a real time human computer interaction (HCI) system when using inhomogeneous backgrounds.

In the next part of our experiments, we are using a Bumblebee2 stereo camera system by Point Grey Research [14] which can calculate the stereo data, i.e. disparity calculation, within hardware and provides images with a resolution of 320×240 with up to 30fps. However, it is not possible to work with higher resolutions values e.g. 1024×768 in real-time. In comparison to the other camera types we achieved the best results with a mean ACC of 99.14% with a mean TPR of 78.24% and FRP of 0.267%

here. The improvement of the stereo cameras is the depth information (Fig. 5d). Thereby the high recognition rate of skin colored pixels results from fading out the background information (Fig. 5c), which is only one of the advantages of stereo camera systems, described in section 2.1. The third kind of camera was an uncooled mobile thermal camera system (FLIR SC600) with a pixel resolution of 640×480 and 30 fps. The camera specifications are: spectral sensibility of 7.5-13.5 µm and thermal sensibility of <45 mK. In figure 6 the ROC curve is graphically presented for different thresholds from 33.6°C to 31.4°C by steps of 0.2°C. For segmentation, the maximal temperature was chosen as a threshold value.



(a) Original (b) Single camera (c) Stereo camera

(d) Depth information (e) Thermal camera

Figure 5. A comparison of segmentation results of different camera systems. (a) the original captured image, (b) the analyzed image without depth information (d), (c) the analyzed image by using depth information and (e) the image from a thermal camera

In our experiments we achieved an average ACC rate of ≈96%. These are not optimal results, because clothes can receive the body temperature and are partially warmer than extremities where the blood pressure value is lower (e.g. hands) as shown in figure 5e. If the face should be segmented only, thermal cameras achieve good results. Because under normal conditions the face has a high blood pressure value and owns within a high body temperature. However, normally the background can also be ignored by thermal cameras. An advantage of thermal cameras is that they can be used in darkness or under low illumination conditions where skin color models are not suitable.

The acquired image segments will now provide the database for further analysis to realize a natural human computer interaction system by using Hidden Markov Model (HMM).

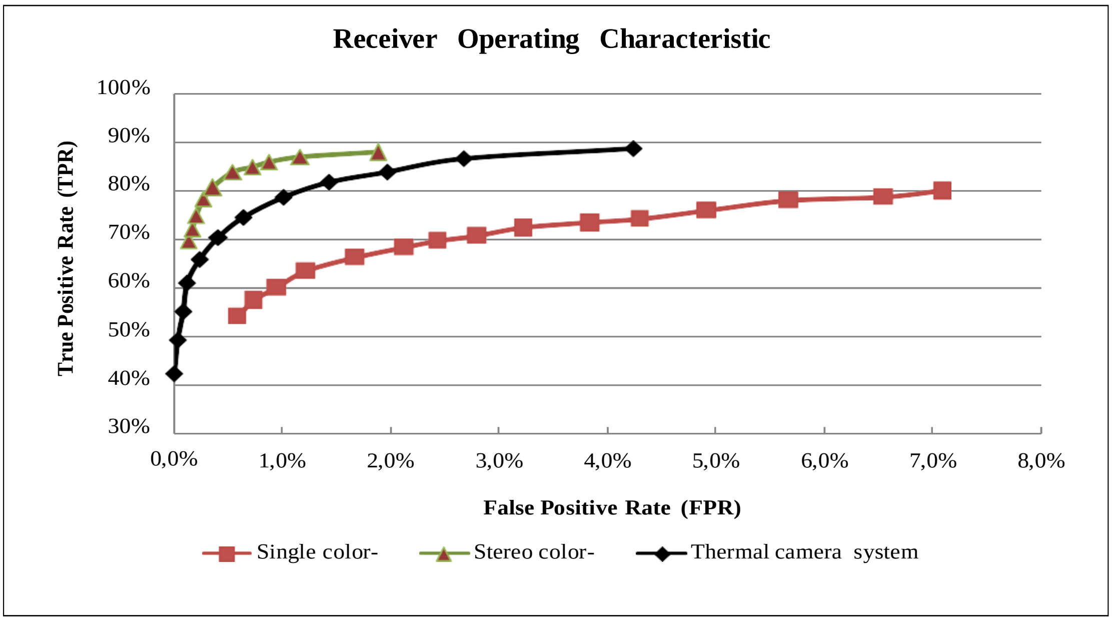


Figure 6. ROC curves for a single color-, stereo color-, and thermal camera system by using different thresholds.

2.4. Feature extraction

There is no doubt about the fact that significant features play a major role in the detection of hand gesticulation. And these features must be independent of person and location in order to realize the natural HCI system. In this manner location, orientation and velocity are the main basic features which can be extracted from hand trajectory. Our previous work showed that using the orientation information as the input data to our system gives the best results in terms of accuracy. Hence, in this work, the angle θ_t was selected between two consequent points of the gesture path which consists of the center of gravity from the respective hand ($x_c(t)$, $y_c(t)$) or the location of fingertips.

$$\theta_t = \arctan\left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t}\right); \; t = 1, 2, \ldots, T - 1 \qquad (7)$$

where T represents the length of the gesture path. For an optimal characterization of motion trajectory, the angle θ_t was quantized in steps of 20° to generate a codeword from 1 to 18. The codeword also included zero code which is specified in. In contrast to other features, an evaluation of the quantized vector is possible at each point of time. This is a major advantage, since it is not necessary to wait till the end of the motion. This discretevector is then used as an input into the classification process.

2.5. Classification

The classification is based on a Hidden Markov Model which is a mathematical model of a stochastic process and it includes three parameters $\Lambda = (\Pi, A, B)$. Thereby Π is defined as the initial vector; A refers to the translation matrix and B represents the emission matrix. These values depend on the construction of HMM which can vary according to the application. For complex applications, a Fully Connected (Ergodic) model can be used, where any state can be reached from other states. In speech recognition, primarily the Left-Right (LR) model is used where any state can change to itself or following states. We have used the Left-Right-Banded (LRB) model in our approach where a state can reach itself or just the next state as shown in figure 7. In order to use HMM three main problems must be solved: Evaluation, Decoding and Training. These problems can be overcome by using Forward-Backward algorithm, Viterbi algorithm and Baum-Welch algorithm respectively .

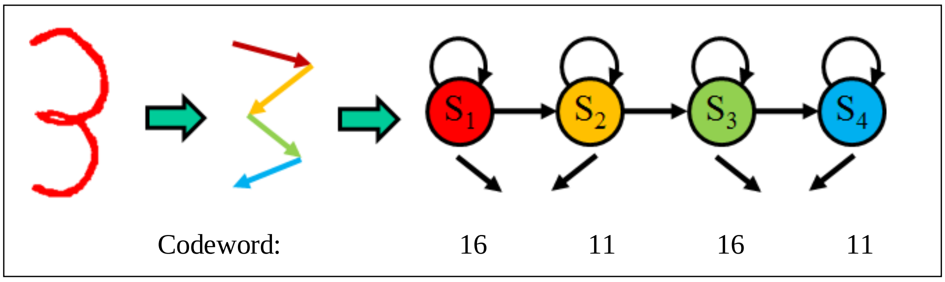


Figure 7. Example of generating the Codeword by LRB model with 4 states.

The isolated and continuous gesture paths are recognized by their discrete vectors using the Forward algorithm to calculate the probability of the best Viterbi path. Moreover, Baum-Welch algorithm is used to do a full training for the initialized HMM parameters to construct a gesture database. The number of states, an example can be seen in figure 7, is based on the complexity of each gesture and is determined by mapping each straight-line segment into a single HMM state. More details about HMM can be found in .

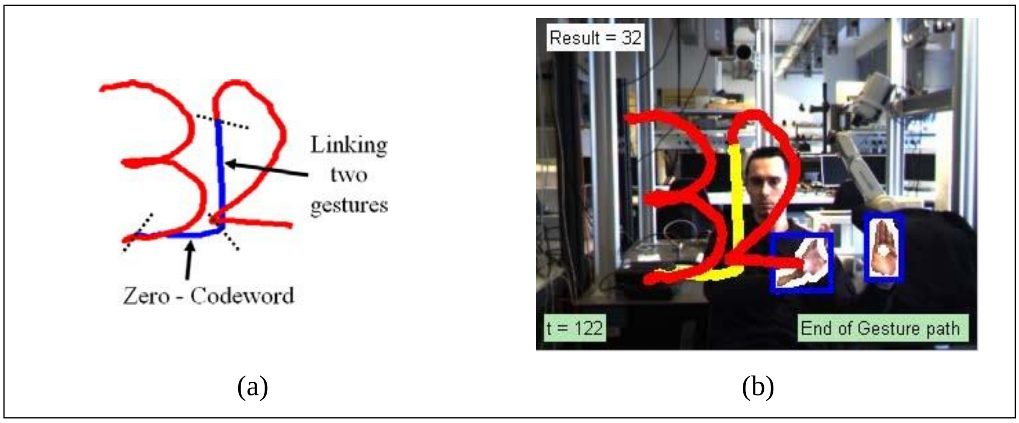


Figure 8. Gesture path of a continuous gesture by using the Zero-Codeword.

In our system, isolated gestures can be recognized from continuous gestures by using the Zero-codeword. Each gesture ends with a line segment, which is assigned to a Zero-codeword. There are many gestures (e.g. E, Z and 2) which contain Zero-codewords in some segmented parts and produce separation problems. To overcome these problems, we assign continuous velocity as a threshold. After detecting the Zero-codeword, a small part of linkage between two gestures must be ignored as shown in figure 8a.

3. Experimental results

In the first part of our proposed system, we describe the classification of 36 isolated gestures (A-Z and 0-9) from stereo color image sequences or online experiments. After detecting the hands and the face by using a skin color model, the motion trajectory is generated and afterwards analyzed by HMM. In our previous work, we designed different types of HMM topologies for comparing and getting the best results by using LRB topology [1]. In our experimental results, each isolated gesture is based on at least 30 video sequences; 20 video samples for training and at least 10 video samples and additionally a lot of online experiments for testing. We achieved an average recognition rate of 98% for isolated gestures by using the LRB topology with 9 states. Figure 9 shows the results of the isolated gesture 'W' at different times. At t=20, the highest probability for getting the gesture 'V' and at t=34, the gesture 'h' are calculated by the system. Finally at t=42 the gesture 'W' was recognized.

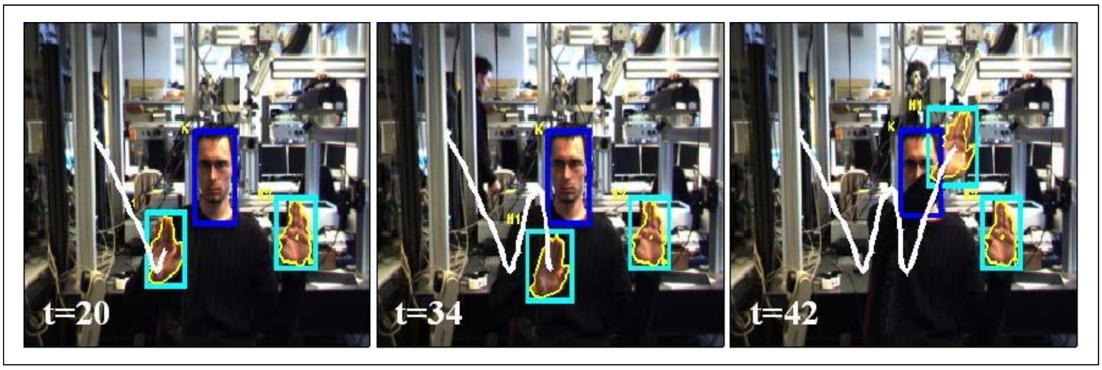


Figure 9. Result of isolated gesture ‘W’ at different times.

The second part is the classification of continuous gestures. For the separation into isolated gestures a Zero-codeword detection is realized by using a constant velocity as threshold. The system has been tested on 70 video sequences for continuous gestures with more than one isolated gesture and a recognition rate of 95.7% is achieved. The system output for the continuous gesture of '90' is shown in figure 10. At t=51 the first gesture is ended with the output '9'. And the linkage between the two gestures is shown at t=68. The second gesture is ended with the output '0' at t=129 and the final result is related to '90'. The input images are captured by Bumblebee2 stereo camera system with 15 fps and 320×240 pixel image resolution. This real-time system was implemented in C++ language.

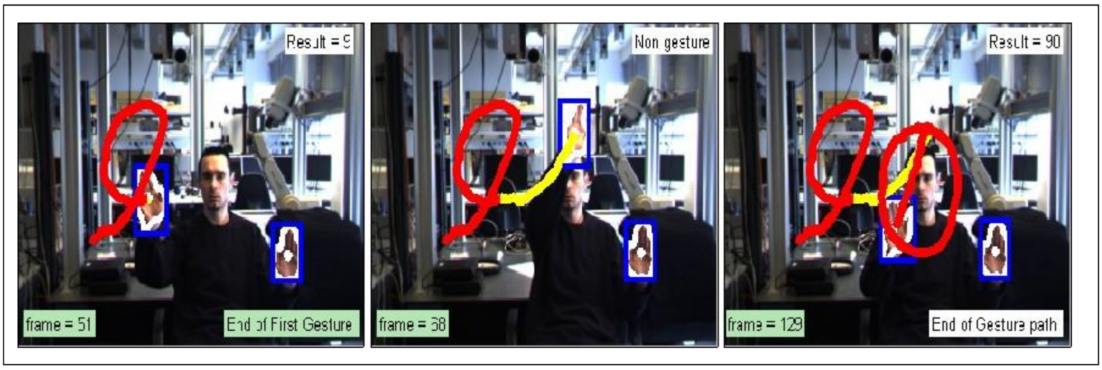


Figure 10. Result of continuous gesture ‘90’ at different time instants.

4. Summary and conclusion

In this paper, we propose an automatic system that recognizes continuous gestures (0-9 & A-Z). Thereby three different types of cameras (single color-, stereo color- and thermal camera) are compared in the area of segmentation. According to our analyses of the advantages and disadvantages and comparison of the receiver operating characteristic (ROC) curves, overall, stereo cameras give the best results for segmentation. Furthermore, it depends on our novel idea of Zero-codeword detection. As a result, the developed system can be used in real-time and allows the user to act in front of the camera online without the requirement of a second person. Our database includes more than 30 video sequences for each gesture. We have achieved an average recognition rate of 98% for isolated gestures and a recognition rate of 95.7% for continuous gestures by using a stereo camera system for segmentation and in addition to that, we have accomplished a lot of online tests. Our future research focuses on the area of recognition from trajectory of the fingertip instead of the center of the hand using a multi camera system.

5. Acknowledgments

This work was supported by Transregional Collaborative Research Centre SFB/TRR 62 ("Companion-Technology for Cognitive Technical Systems") funded by the German Research Foundation (DFG