班级 031111

学号 03111047

本科毕业设计（论文）

外文资料翻译

毕业设计题目 手机拍照手势识别算法

外文资料题目 Hand Gesture Recognition Based on Combined Features Extraction

学　　院 计算机学院

专　　业 计算机科学与技术

学 生 姓 名 吕　莹

指导教师姓名 郭杏莉

World Academy of Science, Engineering and Technology
Vol:3 2009-12-21

Hand Gesture Recognition Based on Combined Features Extraction

Mahmoud Elmezain, Ayoub Al-Hamadi, and Bernd Michaelis
Institute for Electronics, Signal Processing and Communications (IESK)
Otto-von-Guericke-University Magdeburg
{Mahmoud.Elmezain, Ayoub.Al-Hamadi}@ovgu.de

***Abstract*—**Hand gesture is an active area of research in the vision community, mainly for the purpose of sign language recognition and Human Computer Interaction. In this paper, we propose a system to recognize alphabet characters (A-Z) and numbers (0-9) in real-time from stereo color image sequences using Hidden Markov Models (HMMs). Our system is based on three main stages; automatic segmentation and preprocessing of the hand regions, feature extraction and classification. In automatic segmentation and preprocessing stage, color and 3D depth map are used to detect hands where the hand trajectory will take place in further step using Mean-shift algorithm and Kalman filter. In the feature extraction stage, 3D combined features of location, orientation and velocity with respected to Cartesian systems are used. And then, k-means clustering is employed for HMMs codeword. The final stage so-called classification, Baum-Welch algorithm is used to do a full train for HMMs parameters. The gesture of alphabets and numbers is recognized using Left-Right Banded model in conjunction with Viterbi algorithm. Experimental results demonstrate that, our system can successfully recognize hand gestures with 98.33% recognition rate.

***Keywords*—**Gesture Recognition, Computer Vision & Image Processing, Pattern Recognition.

I. Introduction

Sign language recognition from hand motion or hand posture is an active area in gesture recognition research for Human Computer Interaction (HCI). A gesture is spatio-temporal pattern, which may be static or dynamic or both. Static morphs of the hands are called postures and hand movements are called gestures. The goal of gesture interpretation is to push the advanced human-machine communication to bring the performance of human-machine interaction close to human-human interaction. This is due to the existing complexities in hand tracking such as hand appearance, illumination variation, and inter-hands occlusion. These issues undermine the performance and efficiency of tracking algorithms. In the last decade, several methods of potential applications [1], [2], [3], [4], [5] in the advanced gesture interfaces for HCI have been suggested but these differ from one to another in their models. Some of these models are Neural Network [1], HMMs [2] and Fuzzy Systems [5].

Liu and Lovell [6] introduced a system for hand tracking in real-time based on the Camshift algorithm and the compound constant-acceleration Kalman filter algorithm. Whereas Nobuhiko *et al.* [7] used HSV color space to track hands and face in non complex background, where the overlapping between hands and face is solved by matching templates of the previous hands and face. Comaniciu *et al.* [8] proposed a technique to track the moving objects from a moving camera using Mean-shift algorithm and Kalman filter, where the implementation of this technique achieved real-time performance. Mostly, previous approaches have not been considered many points as the combination of accurate segmentation of both hands, robust tracking that containing overlap between hands and face, and the capability of the system to run in real-time on high resolution.

Vassilia et al. [9] developed a system that could recognize both isolated and continuous Greek Sign Language (GSL) sentences where the orientation vector is extracted from images and then used in sentences as input to HMMs. Ho-Sub et al. [10] introduced a hand gesture recognition method, which used the combined features of location, angle and velocity to determine the discrete vector that is used as input to HMMs. This method runs over the alphabets (A-Z), numbers (0-9), six edit commands and six drawing elements. Nianjun et al. [11] proposed a method to recognize the 26 letters from A to Z by using different HMMs topologies with different states number. But, these methods run off-line over a non complex background.

Nguyen et al. [12] proposed a system for hand gesture recognition, where the hand is tracked by Kalman filter and hand blobs analysis to obtain motion descriptors for hand region. This system is fairy robust to background cluster and uses skin color for hand gesture tracking and recognition. Also, the system was tested to a vocabulary of 36 gestures including the American Sign Language letter spelling alphabets and digits. But, this method [12] studies the posture of the hand, not the hand motion trajectory as it is in our system. One of such problems, which arise in real-time hand gesture recognition, is to caused by the fact that the same gesture varies in shape, trajectory and duration, even for the same person. So, HMMs is used in our system where it is capable of modeling spatio-temporal time series.

The main contribution of this paper is to examine the capabilities of combined features of location, orientation and velocity for gesture recognition, which are obtained from spatio-temporal hand gesture path. Additionally, it proposes a real-time capable system that recognize the alphabets characters (A-Z) and numbers (0-9) from stereo color image sequences by the motion trajectory of a single hand using HMMs. Color and 3D depth map are used to detect hands. Furthermore, the hand trajectory is estimated using Mean-shift algorithm [13] and Kalman filter [14] in conjunction with 3D

depth map. The blob segmentation of the hands and face with complex background takes place using 3D depth map from a passive stereo camera, Gaussian Mixture Models (GMM) and color information, which is more robust to the disadvantageous lighting and partial occlusion. The depth information solve the overlapping problem between hands and face. The system is tested on a different experiments with varying features that are extracted from Cartesian systems to decide which feature is the best in terms of result. Each alphabet and each number is based on 30 video (20 for training and 10 for testing). The recognition rate that achieved on testing gestures is 98.33%. The rest of this paper is organized as follow; Section II reviews the basic HMMs technique. Section III demonstrates the suggested system in three subsections. Experimental results are described in Section IV. Finally, the summary and conclusion are presented in Section V.

II. Hidden Markov Models

Markov model is a mathematical model of stochastic process where these processes generate a random sequence of outcomes according to certain probabilities [11], [15], [16], [17], [18]. An HMM is a triple $\lambda = (A, B, \Pi)$ as follows:

- The set of states $S = \{s_1, s_2, ..., s_N\}$ where *N* is a number of states.
- An initial probability for each state Π_i, i=1, 2, ..., *N* such that $\Pi_i = P(s_i)$ at the initial step.
- An *N*-by-*N* transition matrix $A = \{a_{ij}\}$ where a_{ij} is the probability of a transition from state S_i to S_j; $1 \leq i, j \leq N$ and the sum of the entries in each row of matrix *A* must be 1 because this is the sum of the probabilities of making a transition from a given state to each of the other states.
- The set of possible emission (an observation) $O = \{o_1, o_2, ..., o_T\}$ where *T* is the length of gesture path.
- The set of discrete symbols $V = \{v_1, v_2, ..., v_M\}$ where *M* represents the number of discrete symbols.
- An *N*-by-*M* observation matrix $B = \{b_{im}\}$ where b_{im} gives the probability of emitting symbol v_m from state s_i and the sum of the entries in each row of matrix *B* must be 1 for the same pervious reason.

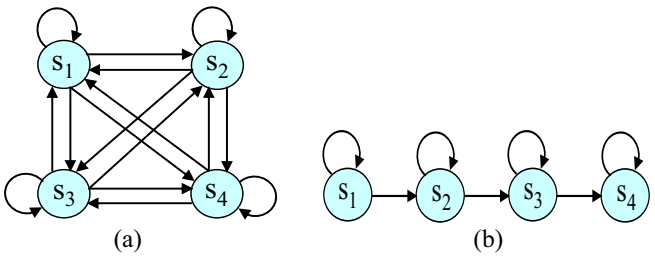


Fig. 1. HMMs topologies with 4 states (a) Ergodic topology (b) LRB topology.

There are three main problems for HMMs: Evaluation, Decoding and Training that can be solved by using Forward or Backward algorithm, Viterbi algorithm and Baum-Welch algorithm respectively [15]. Also, HMMs have a three topology: Fully Connected (Ergodic model) where any state in it can be reached from other states, Left-Right model such that each state can go back to itself or to the following states and Left-Right Banded (LRB) model that also each state can go back to itself or the following state only (Fig. 1).

III. Suggested System

We propose an automatic system that recognizes isolated gesture for Alphabets (A-Z) and numbers (0-9) in real-time from stereo color image sequences by the motion trajectory of a single hand using HMMs. In particular, the proposed system consists of three main stages; an automatic hand segmentation and tracking, feature extraction and classification (Fig.2).

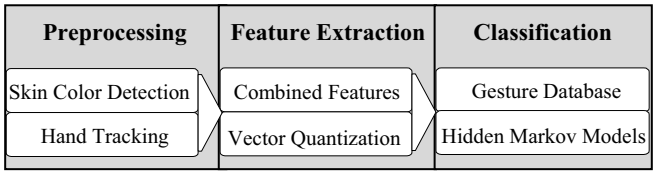


Fig. 2. Simplified structure showing the main computational modules for gesture recognition system.

- **Preprocessing**; localize and track the hand to generate its motion trajectory (gesture path).
- **Feature extraction**; Clustering extracted features to generate discrete vectors, which are used as input to HMMs recognizer.
- **Classification**; the gesture path is recognized using discrete vector and Left-Right Banded topology.

A. Automatic Hand Segmentation and Tracking

In this paper, a method for detection and segmentation of a hand in stereo color images with complex background is described where the hand segmentation takes place using 3D depth map and color information. Segmentation of skin colored regions becomes robust if only the chrominance is used in analysis. Therefore, YC_bC_r color space is used in our system where *Y* channel represents brightness and (C_b, C_r) channels refer to chrominance. We ignore *Y* channel to reduce the effect of brightness variation and use only the chrominance channels that fully represent the color information. A large database of skin and non-skin pixels is used to train the Gaussian model. The Gaussian Mixture Model begins with modeling of skin using skin database where a variant of k-means clustering algorithm [17], [19], [20] performs the model training to determine the initial configuration of GMM parameters. For the skin segmentation of hands and face in stereo color image sequences an algorithm is used, which calculates the depth value in addition to skin color information. The depth information (Fig. 3(c)) solves the overlapping problem between hands and face where it is obtained by passive stereo measuring based on cross correlation and the known calibration data of the cameras. Several clusters are composed of the resulting 3D-points. The clustering algorithm can be considered as kind of region growing in 3D that used two criteria; skin color and Euclidean distance. Furthermore, this method is more robust to the disadvantageous lighting and partial occlusion, which occur in real-time environment. In addition, blob analysis is used to derive the boundary area,

bounding box and hands centroid point. For more details, the reader can refers to [2], [21].

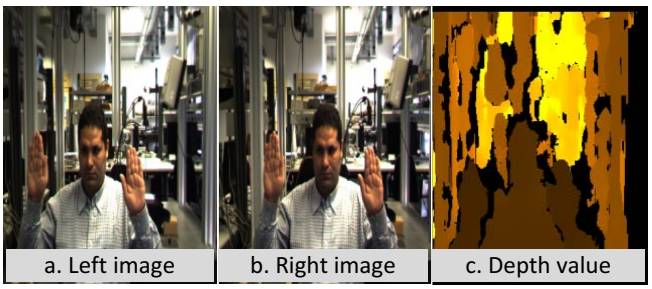


Fig. 3. (a) Left image frame of video stream. (b) Right image. (c) The depth value of left and right image via the Bumblebee stereo camera.

After localization of the hand's target from the segmentation step, we find its color histogram with Epanechnikov kernel [8]. This kernel assigns smaller weights to pixels father from the center to increases the robustness of the density estimation. To find the best match of our hand target in the sequential frames, the Bhattacharyya coefficient [22] is used to measure the similarity by maximizing Bayes error that arising from the comparison of the hand target and candidate. We take in our consideration the mean depth value that is computed from the previous frame for the hand region to solve the overlapping between hands and face. The mean-shift procedure is defined recursively and performs the optimization to compute the mean shift vector. After each mean-shift optimization that gives the measured location of the hand target, the uncertainty of the estimate can also be computed and then followed by the Kalman iteration, which drives the predicated position of the hand target. Thereby, the hand gesture path is obtained by taking the correspondences of detected hand between successive image frames (Fig. 5(d)). For more details, the reader can refer to [2], [8], [21].

B. Feature Extraction

There is no doubt that selecting good features to recognize the hand gesture path plays significant role in system performance. There are three basic features; location, orientation and velocity. We analyze the effectiveness of these features that are extracted from a hand trajectory and also combine them to test their recognition rate. A gesture path is spatio-temporal pattern that consists of hand centroid points (x_{hand}, y_{hand}). The coordinates in the Cartesian space can be extracted from gesture frames directly. We consider two types of location features. The first location feature is Lc that measures the distance from the centroid point to all points of the gesture path because different location features are generated for the same gesture according to different starting points (Eq.1). The second location feature is Lsc , which is computed from the start point to the current point of hand gesture path (Eq.3).

$$Lc_t = \sqrt{(x_{t+1} - C_x)^2 + (y_{t+1} - C_y)^2} \quad (1)$$

$$(C_x, C_y) = \frac{1}{n}(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n} y_t) \quad (2)$$

$$Lsc_t = \sqrt{(x_{t+1} - x_1)^2 + (y_{t+1} - y_1)^2} \quad (3)$$

where $t = 1, 2, ..., T-1$ and T represents the length of hand gesture path. (C_x, C_y) refers to the centroid of gravity at the n points. To verify the real-time implementation, the centroid point of gesture path is computed after each frame.

The second basic feature is the orientation, which gives the direction along that the hand when traverses in space during the gesture making process. As described above, the orientation feature is based on the calculation of the hand displacement vector at every point and is represented by the orientation according to the centroid of gesture path (θ_{1t}), the orientation between two consecutive points (θ_{2t}) and the orientation between start and current gesture point (θ_{3t}).

$$\theta_{1t} = \arctan\left(\frac{y_{t+1} - C_y}{x_{t+1} - C_x}\right) \quad (4)$$

$$\theta_{2t} = \arctan\left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t}\right) \quad (5)$$

$$\theta_{3t} = \arctan\left(\frac{y_{t+1} - y_1}{x_{t+1} - x_1}\right) \quad (6)$$

The third basic feature is the velocity, which plays an important role during gesture recognition phase particulary at some critical situations. The velocity is based on the fact that each gesture is made at different speeds where the velocity of the hand decreases at the corner point of a gesture path. The velocity is calculated as the Euclidean distance between the two successive points divided by the time in terms of the number of video frames as follows;

$$V_t = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2} \quad (7)$$

In the Cartesian coordinate system, we use different combination of features to obtain a variety feature vectors. For example, the feature vector at frame $t+1$ is obtained by union of locations features (Lc_t, Lsc_t), locations features with velocity feature (Lc_t, Lsc_t, V_t), orientations features $(\theta_{1t}, \theta_{2t}, \theta_{3t})$, orientations features with velocity feature $(\theta_{1t}, \theta_{2t}, \theta_{3t}, V_t)$ and locations features with orientations features and velocity feature $(Lc_t, Lsc_t, \theta_{1t}, \theta_{2t}, \theta_{3t}, V_t)$.

Each frame contains a set of feature vectors at time t where the dimension of space is proportional to the size of feature vectors. In this manner, gesture is represented as an ordered sequence of feature vectors, which are projected and clustered in space dimension to obtain discrete codeword that are used as an input to HMMs. This is done using k-means clustering algorithm [19], [20], which classifies the gesture pattern into K clusters in the feature space.

1) Vector Quantization: The extracted features are quantized to obtain the discrete symbols. When the basic features such as locations and velocity are used separately, these features are normalized and multiplied by a different scalar ranging from 10 to 30. On the other side, the normalization of the orientation features is to divide by 10°, 20°, 30° and 40° to obtained its codeword. In addition to the combination features in the Cartesian system, we use k-mean clustering algorithm to classify the gesture feature into K clusters on the feature space. This algorithm is based on the minimum

distance between the center of each cluster and the feature point. We divide the set of feature vectors into set of clusters. This allows us to model the hand trajectory in the feature space by one cluster. The calculated cluster index is used as input (i.e. observation symbol) to the HMMs. Furthermore, we usually do not know the best number of clusters in the data set. In order to specify the number of clusters K for each execution of the k-means algorithm, we considered $K = 28, 29, ..., 37$, which is based on the numbers of segmented parts in all alphabets character (A-Z) and numbers (0-9) where each straight-line segment is classified into single cluster.

Suppose we have n sample of trained feature vectors $x_1, x_2, ..., x_n$ all from the same class, and we know that they fall into k compact clusters, $k < n$. Let m_i be the mean of the vectors in cluster i. If the clusters are well separated, a minimum distance classifier is used to separate them. That is, we can say that x is in cluster i if $\|x - m_i\|$ is the minimum of all the k distances. The following procedure shows that how to find the k-means;

- Build up randomly an initial Vector Quantization Codebook for the means $m_1, m_2, ..., m_k$
- Until there are no changes in any mean
  - Use the estimated means to classify each sample of train vectors into one of the clusters m_i
    for i=1 to k
    Replace m_i with the mean of all of the samples of trained vector for cluster i
    end (for)
- end (Until)

A general observation is that different gestures have different trajectories in the cluster space, while the same gesture show very similar trajectories.

C. Classification

The final stage in our system is classification. Throughout this stage, Baum-Welch algorithm [15] is used to do a full training for the initialized HMMs parameters to construct gestures database. Each reference pattern in the gesture database for alphabets (A-Z) and numbers (0-9) is modeled by Left-Right Banded model with varying number of states ranging from 3 to 6 states based on its complexity. As, the excessive number of states can generate the over-fitting problem if the number of training samples is insufficient compared to the model parameters. The hand gesture path is classified by selecting the maximal observation probability of gestures model. The maximal gesture model is the gesture whose observation probability is the largest among all 36 gestures (A-Z & 0-9). The type of observed gesture (O) is decided by Viterbi algorithm frame by frame (i.e. accumulatively until it receives the gesture end signal). The following steps show how the Viterbi algorithm works on gesture model $\lambda_g(a^g, b^g, \Pi^g)$ (Fig. 4):

1. Initialization: $for\ 1 \le i \le N$,
   - $\delta_1^g(i) = \Pi_i^g . b_i^g(o_1)$
2. Recursion (accumulative observation probability computation): $for\ 2 \le t \le T,\ 1 \le j \le N$,
   - $\delta_t^g(j) = \max_i[\delta_{t-1}^g(i).a_{ij}^g].b_j^g(o_t)$
3. Termination:
   - $P(O|\lambda_g) = \max_i[\delta_T^g(i)]$

where N is the number of states, Π_i^g represents the initial value for the state i, a_{ij}^g is the transition probability from state i to state j, $b_j^g(o_t)$ refers to the probability of emitting o at time t in state j, and $\delta_t^g(j)$ represents the maximum likelihood value in state j at time t.

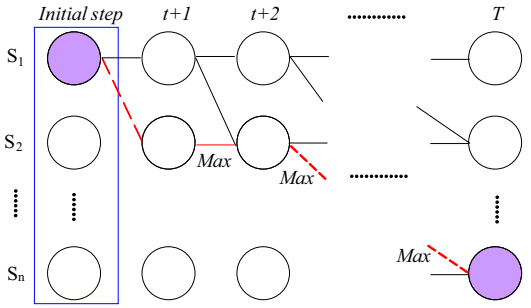


Fig. 4. The best path for LRB model with N states where it starts from S_1 to S_N, N=3, 4, ..., 6 and t=1.

IV. Experimental Results

Our proposed system has been tested on various video sequences with various hand shape as well as overlapping and partial occlusion. Our proposed system was capable for real-time implementation and showed good results to recognize alphabets character and numbers from stereo color image sequences via the motion trajectory of a single hand using HMMs. The input images were captured by Bumblebee stereo camera system that has 6 mm focal length at 15FPS with 240×320 pixels image resolution, Matlab implementation. In our experimental results, each isolated gesture was based on 30 video sequences, which 20 video samples for training by BW algorithm and 10 video samples for testing (Totally, our database contains 720 video sample for training and 360 video sample for testing). The gesture recognition module match the hand gesture path against the database of reference gestures, to classify which class it belongs to. The higher priority was computed by Viterbi algorithm to recognize the alphabets and numbers in real-time frame by frame over LRB topology with different number of states ranging from 3 to 6.

We test the importance of the three basic features (location, orientation, velocity) in the Cartesian coordinate. Moreover, the observation sequence for HMMs is quantified either by normalization in case of separated features or by the k-means clustering algorithm in case of combined features. From table 1, the recognition ratio of isolated gestures achieved best results with 98.33% using $(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$ feature. The recognition ratio is the number of correctly recognized gestures to the number of tested gestures (Eq. 8).

$$Reco.\ ratio = \frac{\#\ recognized\ gestures}{\#\ test\ gestures} \times 100\% \quad (8)$$

According to the separated features in Fig. 5(a) & (b), the orientation features $(\theta_1, \theta_2, \theta_3)$ are better rather than the

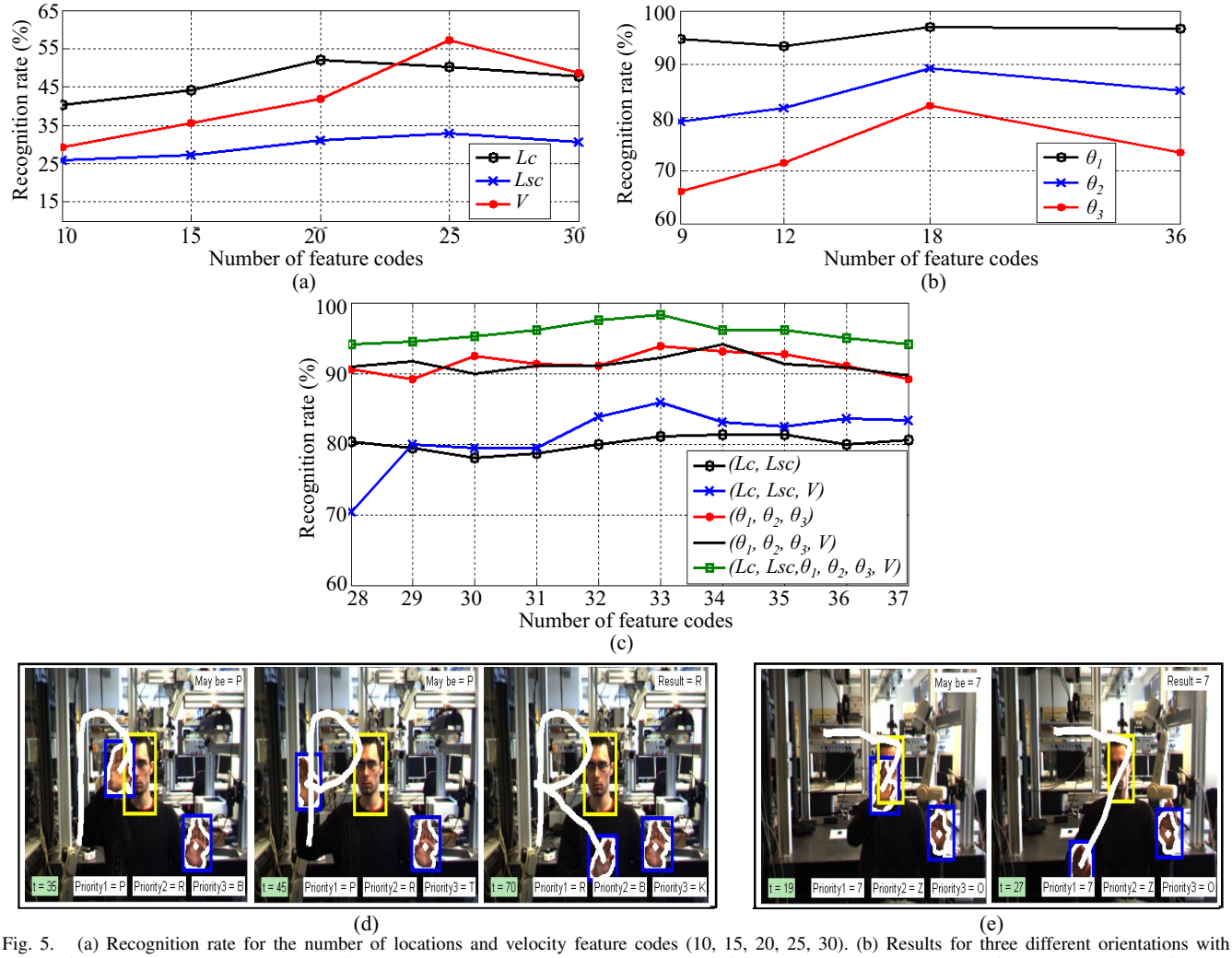


Fig. 5. (a) Recognition rate for the number of locations and velocity feature codes (10, 15, 20, 25, 30). (b) Results for three different orientations with varying feature codes number (9, 12, 18, 36). (c) Recognition rate according to a combined features in Cartesian system over feature codes number from 28 to 37. (d) The high priority is alphabet 'P' at t=45 and at t=70 the result is 'R'. (e) Solving overlap problem between hand and face at t=19 and the high priority is '7' at t=27.

recognition rate of the locations (Lc, Lsc) or the velocity (V) features. This in turn leads to the orientation feature $(\theta_1 = 96.94\%)$ is the most effective among the three basic features. Furthermore, the velocity feature shows a lower discrimination power (57.22%) than that of the orientation features. Also, the Lsc feature result is the lowest recognition rate of 32.78%. Additionally, the (Lc, Lsc, V), $(\theta_1, \theta_2, \theta_3, V)$ and $(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$ features, which include the velocity information show higher recognition than when using the velocity information (Fig. 5(c)). In short, Fig. 5 show the results of the experiments that were performed to determine the best feature codes number (the best number of feature code is 33 for the feature $(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$). Fig. 5(d) & (e) shows the output of the system for gesture alphabet 'R' and number '7' respectively, in addition to the solved overlapping problem between hand and face by 3D depth map.

V. Conclusion

This paper proposes a system to recognize the alphabets character (A - Z) and numbers (0 - 9) from stereo color image sequences by the motion trajectory of a single hand using HMMs. This system uses the combined features of location, orientation and velocity for Cartesian systems. We have shown that the effective of these features can yield reasonable recognition rates. The database contains 720 video samples for training and 360 video sequences for testing the isolated gestures. The proposed system has shown good performance when applied on several video samples containing confusing situations such as partial occlusion and overlapping. The results show that; the proposed system is suitable for real-time application and can successfully recognize hand gestures with 98.33% recognition rate. The future research will address the hand gesture spotting and recognition for a sentence using fingertip instead of the hand centroid point in conjunction with multi-camera system.

Acknowledgment

This work is supported by BMBF Bernstein-Group (FKZ: 01GQ0702), LSA grants (C4-NIMITEK 2, FKZ: UC4-3704M), Forschungspraemie (BMBF-Frderung, FKZ: 03FPB00213) and Transregional Collaborative Research Centre SFB/TRR 62 "Companion-Technology for Cognitive Technical Systems" funded by DFG.

TABLE I
RESULTS OF HAND GESTURES ACCORDING TO DIFFERENT FEATURES EXTRACTION IN CARTESIAN SYSTEM WITH THE BEST FEATURE CODE NUMBER.

Feature type	Feature space	# Feature code	Training data	Hand gestures results		
				Testing data	Correct data	Recognition (%)
Separated in Cartesian coordinates	Lc	20	720	360	187	51.94
	Lsc	25	720	360	118	32.78
	v	25	720	360	206	57.22
	θ_1	18 ; 36	720	360	349	96.94
	θ_2	18	720	360	321	89.17
	θ_3	18	720	360	236	82.22
Union in Cartesian coordinates	(Lc, Lsc)	35	720	360	293	81.39
	(Lc, Lsc, V)	33	720	360	309	85.83
	$(\theta_1, \theta_2, \theta_3)$	33	720	360	338	93.89
	$(\theta_1, \theta_2, \theta_3, V)$	34	720	360	339	94.16
	$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$	33	720	360	354	98.33

REFERENCES

[1] X. Deyou, *A Network Approach for Hand Gesture Recognition in Virtual Reality Driving Training System of SPG*, International Conference ICPR, pp. 519-522, 2006.
[2] M. Elmezain, A. Al-Hamadi, and B. Michaelis, *Real-Time Capable System for Hand Gesture Recognition Using Hidden Markov Models in Stereo Color Image Sequences*, The Journal of WSCG, Vol. 16, No. 1, pp. 65-72, 2008.
[3] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, *A Hidden Markov Model-Based Continuous Gesture Recognition System for Hand Motion Trajectory*, International Conference on Pattern Recognition (ICPR) pp. 1-4, 2008.
[4] M. Elmezain, A. Al-Hamadi, and B. Michaelis, *A Novel System for Automatic Hand Gesture Spotting and Recognition in Stereo Color Image Sequences*, The Journal of WSCG, Vol. 17, No. 1, pp. 89-96, 2009.
[5] E. Holden, R. Owens, and G. Roy, *Hand Movement Classification Using Adaptive Fuzzy Expert System*, The Journal of Expert Systems, Vol. 9(4), pp. 465-480, 1996.
[6] N. Liu, and B. C. Lovell, *MMX-accelerated Real-Time Hand Tracking System*, In IVCNZ, pp. 381-385, 2001.
[7] T. Nobuhiko, S. Nobutaka, and S. Yoshiaki, *Extraction of Hand Features for Recognition of Sign Language Words*, In International Conference of VI, pp. 391-398, 2002.
[8] D. Comaniciu, V. Ramesh, and P. Meer, *Kernel-Based Object Tracking*, The IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 25, pp. 564-577, 2003.
[9] N. P. Vassilia and G. M. Konstantinos, *On Feature Extraction and Sign Recognition for Greek Sign Language*, International Conference on Artificial Intelligence and Soft Computer, pp. 93-98, 2003.
[10] Y. Ho-Sub, S. Jung, J. B. Young, and S. Y. Hyun, *Hand Gesture Recognition using Combined Features of Location, Angle and Velocity*, Journal of Pattern Recognition, Vol. 34(7), pp. 1491-1501, 2001.
[11] L. Nianjun, C. L. Brian, J. K. Peter, and A. D. Richard *Model Structure Selection & Training Algorithms for a HMM Gesture Recognition System*,International Workshop IWFHR, pp. 100-105, 2004.
[12] D. B. Nguyen, S. Enokida, and E. Toshiaki, *Real-Time Hand Tracking and Gesture Recognition System*, In GVIP Conf., pp. 362-368, 2005.
[13] D. Comaniciu, V. Ramesh, and P. Meer, *Real-Time Tracking of Non-Rigid Objects Using Mean Shift*, In Conference CVPR, pp. 1-8, 2000.
[14] G. Welch, and G. Bishop, *An Introduction to the Kalman Filter*, In Technical Report, University of North Carolina at Chapel Hill, pp. 95-041, 1995.
[15] R. R. Lawrence, *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition*, Proceeding of the IEEE, Vol. 77(2), pp. 257-286, 1989.
[16] S. Mitra, and T. Acharya, *Gesture Recognition: A Survey*, IEEE Transactions on Systems, MAN, and Cybernetics, pp. 311-324, 2007.
[17] M. Elmezain, A. Al-Hamadi, S. S. Pathan, and B. Michaelis, *Spatio-Temporal Feature Extraction-Based Hand Gesture Recognition for Isolated American Sign Language and Arabic Numbers*. IEEE Symposium on ISPA, pp. 254-259, 2009.
[18] M. Elmezain, A. Al-Hamadi, G. Krell, S. El-Etriby, and B. Michaelis, *Gesture Recognition for Alphabets from Hand Motion Trajectory Using Hidden Markov Models*, The IEEE International Symposium on Signal Processing and Information Technology, pp. 1209-1214, 2007.
[19] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, *An Efficient k-means Clustering Algorithm: Analysis and Implementation*, IEEE Transaction on PAMI, Vol. 24, pp. 881-892, 2002.
[20] C. Ding and X. He, *K-means Clustering via Principal Component Analysis*, International Conference on ML, pp. 225-232, 2004.
[21] R. Niese, A. Al-Hamadi, and B. Michaelis, *A Novel Method for 3D Face Detection and Normalization*, Journal of Multimedia, Vol. 2, pp. 1-12, 2007.
[22] S. Khalid, U. Ilyas, S. Sarfaraz, and A. Ajaz, *ABhattacharyya Coefficient in Correlation of Gary-Scale Objects*, The Journal of Multimedia, Vol. 1, pp. 56-61, 2006.

Mahmoud Elmezain was born in Egypt. He received his Masters Degree in Computer Science in 2004. Between 1997 and 2004 he worked as Demonstrator in Dept. of Statistic and Computer Science. Since 2004 he is Assistant lecturer in Dept. of Computer Science, Faculty of Science, Tanta University, Egypt. His current work on a Ph.D. thesis focuses on image processing, pattern recognition and human-computer interaction, at the Institute for Electronics, Signal Processing and Communications at Otto-von-Guericke University of Magdeburg, Germany.

Ayoub K. Al-Hamadi was born in Yemen in 1970. He received his Masters Degree in Electrical Engineering & Information Technology in 1997 and his Ph.D. in Technical Computer Science at the Otto-von-Guericke University of Magdeburg, Germany in 2001. Since 2002 he has been Assistant Professor and 2005 Post-Doc in KFST in Magdeburg. 2004 until 2005 he graduated Professional Training for Industrial Project Management and Start-Up of Business Establishment at University Magdeburg, Germany. Between 2006 and 2008 he has been Junior-Research-Group-Leader at the Institute for Electronics, Signal Processing and Communications at the Otto-von-Guericke University Magdeburg. In August 2008 he became Professor of Neuro-Information Technology at the Otto-von-Guericke University Magdeburg. His research work concentrates on the field of image processing, computer vision, pattern recognition, human-computer interaction, artificial intelligence and information technology. Prof. Dr.-Ing. Al-Hamadi is the author of more than 100 articles in peer-reviewed international journals and conferences.

Bernd Michaelis was born in Magdeburg, Germany in 1947. He received a Masters Degree in Electronic Engineering from the TH Magdeburg in 1971 and his first Ph.D. in 1974. Between 1974 and 1980 he worked at the TH Magdeburg and was granted a second doctoral degree in 1980. In 1993 he became Professor of Technical Computer Science at the Otto-von-Guericke University Magdeburg. His research work concentrates on the field of image processing, artificial neural networks, pattern recognition, processor architectures, and microcomputers. Professor Michaelis is the author of more than 200 articles.

基于组合特征提取技术的手势识别

摘 要

手势在视觉交流上是一个热门的研究领域，主要用于手语识别和人机交互的目的。在本论文中，我们提出了一个通过使用隐马尔可夫模型（HMM 模型）能够实时从彩色图像的序列中识别字母字符（A-Z）和数字（0-9）的系统。我们的系统有三个主要阶段：自动分割和手势区域的预处理、特征提取和分类。在自动分割和手势区域的预处理阶段，通过使用均值漂移算法和卡尔曼滤波，颜色和 3D 深度图是用来探测手将出现的轨迹。在特征提取阶段，笛卡尔系统的使用让我们得到三维组合特征的位置、方向和速度。然后，K-均值聚类采用隐马尔可夫模型。最后阶段所谓的分类，Baum - Welch 算法是用来做一个完整的隐马尔可夫模型参数训练。通过使用左-右手型与 Viterbi 算法结合的方法字母和数字的手势被识别。

实验表明我们的系统能成功识别手势的概率是 98.33%。

关键字：手势识别，计算机视觉，图像处理，模式识别

摘　要

1. 前言

从手势运动和手势位置得到的手语识别是一个用于人机交互的热门研究领域。一个手势是时空格局，这可能是静态或动态或两者。静态的手被称为姿势，而动态的手称作手势。手势解释的目的是推动人机交互从而使人机交互性能接近人际交往。这是由于手势跟踪存在的复杂性，如手的外观，光照变化，跨手闭塞。这些问题损害了跟踪算法的性能和效率。在过去的十年中，几种方法在先进的手势交互的应用前景[1][2][3][4][5]已经被提出来了，但这些差异在他们的模型中又都不相同。这些模型是神经网络[1]，隐马尔可夫模型[2]与模糊系统[5]。

Liu and Lovell 介绍了一种基于 Camshift 实时算法和复合恒定加速度卡尔曼滤波算法的实时手势跟踪系统。而 Nobuhiko 等人用 HSV 颜色空间来追踪非复杂背景下的手和脸，那里的重叠的手和脸通过先前的手和脸模板匹配能较好地分离开。Comaniciu 等人提出了一种使用均值漂移算法和卡尔曼滤波跟踪从摄像头获得的移动物体的技术，这一技术主要获得了实时跟踪性能。先前的技术没有考虑到许多点，如双手的准确分割组合，包括手和脸重叠的鲁棒性跟踪和系统实时高分辨率的运行能力。

Vassilia 等人开发了一种系统，可以识别孤立和连续的希腊手势语言，其中方向向量是从图像中提取，然后作为参数输入到隐马尔可夫模型在句子中被使用。Ho-Sub 等人介绍了手势识别方法，该方法使用位置，角度和速度的组合特征确定作为输入到隐马尔可夫模型的离散向量。这种方法在字母（A-Z），数字（0-9），六编辑命令和六个绘图元素上可以实现。Nianjun 等人通过使用不同的隐马尔可夫模型的拓扑结构数提出一个方法来识别不同国家的从 A 到 Z 的 26 个字母。但是，这些方法运行在一个非复杂离线背景。

Nguyen 等人提出一个手势识别系统，在这个系统里通过卡尔曼滤波和手斑点分析，以获得手部区域动作的描述来跟踪手势。这个系统对背景聚簇和使用皮肤颜色跟踪和识别手势相当强大。此外该系统用包括美国手语拼写字母和数字的 36 个词汇来测试。但是这种方法在我们的系统中研究手的姿势而不是手的运动轨迹。其中有这样一个问题，它提高了手势识别的实时性，是由事实所引起的同样的手势如形状、轨迹和持续时间，甚至是同一个人变化引起的。所以，隐马尔可夫模型是在我们的系统用在它有能力建模时空的时间序列。

本文的主要贡献是研究用于手势识别的位置、方向和速度的组合特征的作用，这个特征是从时空手势路径获得的。此外，它提出了一个能够从三维颜色图像序列中通过使用隐马尔可夫模型模型跟踪单个手势运动轨迹来识别字母字符（A-Z）和数字（0-9）的实时系统。颜色和 3D 深度图是用来检测手。此外，手的轨迹采用均值漂移算法[13]和卡尔曼滤波[14]与 3D 深度图结合的办法来确定。手和脸从来自立体相机、高斯混合模型（GMM）的和颜色信息的三维深度图从复杂背景分割出来，这相对于不利的照明和部分遮挡是更强大的。深度信息解决了手和脸重叠问题。该系统是用来自笛卡尔系统的变化的特征在不同的实验上测试以决定哪个特征能得到最好的结果。每个字母和数字用 30 帧测试（20 帧用来训练和 10 帧用来测试）。测试的手势在识别率上有 98.33%。本文的其余部分如下：

第二部分介绍基本隐马尔可夫模型技术。

第三部分在三个小节中证明这个系统。

第四部分说明实验结果。

最后，第五部分提出总结和结论。

2．隐马尔可夫模型

马尔可夫模型是一个随机过程的数学模型，它在处理过程中产生一个具有相应概率密度分布的状态序列。一个隐马尔可夫模型是由三元组参数$\lambda=(A,B,\prod)$表示如下：

• 一个状态集$S=\{s_1,s_2,...,s_N\}$，N 为常数。

• 一个初始可能每一状态$\prod_i, i=1,2,...,N$，这样第一步$\prod_i=P(s_i)$。

•一个 N 到 N 转移矩阵$A=\{a_{ij}\}$，其中a_{ij}是从状态s_i到s_j的转移的可能性；$1\le i,j\le N$和矩阵 A 的每行之和必须是 1，因为这是让一个给定状态到每一其他状态转移的可能性总和。

•可能的观察序列集$O=\{o_1,o_2,...,o_T\}$，其中 T 是手势路径的长度。

•离散的信号集$V=\{v_1,v_2,...,v_N\}$，其中 M 是离散的信号。

•一个 N 到 M 的观察矩阵$B=\{b_{im}\}$，其中b_{im}给出来至状态s_i的信号v_m的可能值，而矩阵 B 的每行值总和必须是 1，原因和前面的一样。

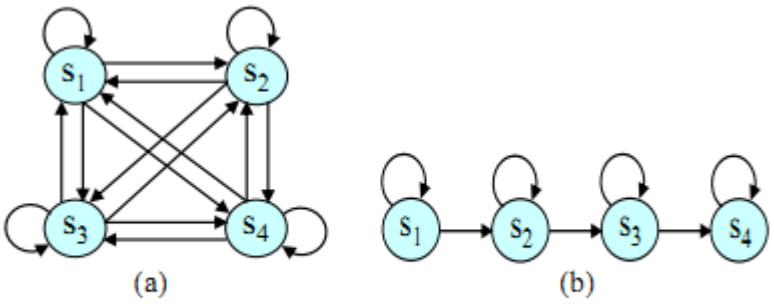


Fig. 1. HMMs topologies with 4 states (a) Ergodic topology (b) LRB topology.

对隐马尔可夫模型来说有三个主要问题：计算问题、解读问题、训练问题。这三个问题可以分别通过前向和后向算法、Viterbi 算法和 Baum- Welch 算法解决。此外，隐马尔可夫模型的拓扑结构有三种：完全连接（遍历模型），在这个结构里可以从任一状态到达其他状态；左-右模型，在这个模型里每个状态只能到达自己状态和下一状态；左-右带状模型，在这个模型里每个状态只能到达自己状态和下一状态（图 1）。

3．推荐系统

我们开发了一个自动识别系统，这个系统可以从三维色彩图片序列中通过单个手的运动轨迹使用隐马尔可夫模型实时识别代表字母（A-Z）和数字（0-9）的手势。特别是，这个系统包括三个主要阶段：手势自动分割和追踪阶段、特征提取阶段和分类阶段（图 2）。

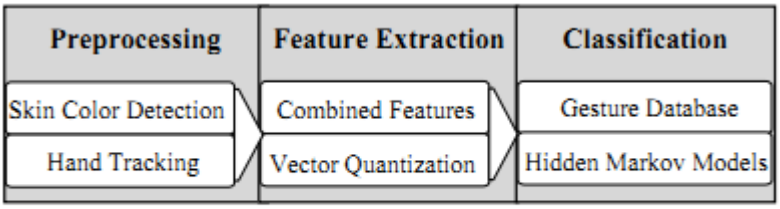


Fig. 2. Simplified structure showing the main computational modules for gesture recognition system.

•预处理：定位和跟踪的手以产生它的运动轨迹（手势路径）。

•特征提取：聚类提取的特征生成作为输入参数到隐马尔可夫模型识别使用的离散向量。

•分类：通过使用离散向量和左-右带状拓扑结构识别手势路径。

3.1 手势自动分割和追踪

本文描述了探测和分割复杂背景下的三维色彩图片里的手势的方法，在这个方法里使用 3D 深度图和色彩信息来分割手势。皮肤颜色区域的分割只有在将色度应用在分析中才能变强大。所以，在我们的系统中使用 YCbCr 颜色空间，其中 Y 分量代表亮度，而（Cb，Cr）分量是指色度。我们忽略 Y 分量，以减少亮度变化的影响，仅使用色度分量，这样充分得到颜色信息。一个大型肤色和非肤色像素数据库被用来训练高斯模型。高斯混合模型使用皮肤数据库开始建立皮肤模型，其中大量 k-均值聚类算法用来模型的训练，以确定 GMM 参数的初始配置。对于三位色彩图像序列中的手和脸的肤色分割，我们计算肤色深度值以增加肤色信息。深度信息（图 3（c））解决了由基于相互关联交叉和已知坐标的照相机位置数据而测量获得的手和脸的重叠问题。几组值组成了最终的三维坐标点。聚类算法可以看成是种在三维空间中的使用了两种准则的区域生长，这两种准则是：皮肤颜色和欧氏距离。此外，这种方法对于实时环境中发生的不利照明和部分遮挡有更好

的鲁棒性。还有，点分析被用来推测边界面积、周长和形心。欲了解更多详情，读者可以阅读[2]，[21]。

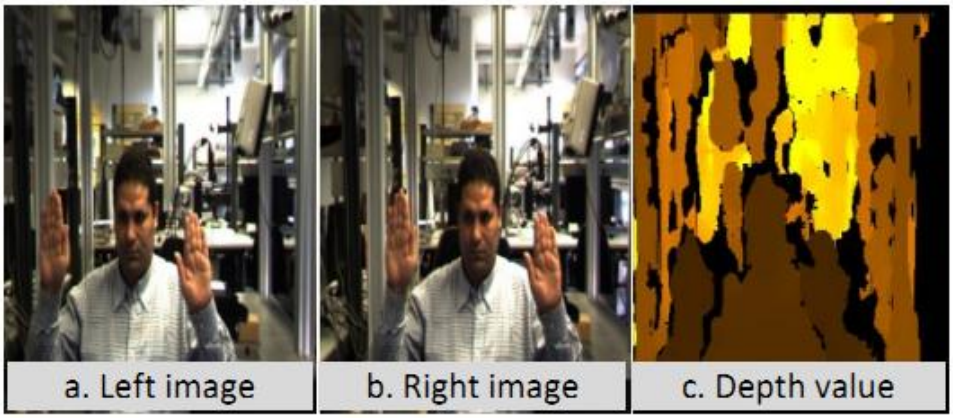


Fig. 3. (a) Left image frame of video stream. (b) Right image. (c) The depth value of left and right image via the Bumblebee stereo camera.

经过来至分割步骤的手势目标定位后，我们发现手的颜色直方图和 Epanechnikov 核。这个核分配来至中心的父像素较小的权重以增加密度估计的鲁棒性。为了找到在连续帧中手势目标的最佳匹配，我们使用 Bhattacharyya 系数通过从手势目标和模板的比较得到的贝叶斯误差来测量相似程度。我们对先前的帧计算手势区域的平均深度值以解决手和脸的重叠问题。平均偏移过程被定义为递归地和执行优化计算平均偏移向量。经过每个均值偏移优化，系统给出手势目标的测量位置。能够计算不确定的估计值，然后通过卡尔曼迭代迭代得到手势目标。因此，我们可以通过检测连续的图像帧之间的手的关系得到手势路径（图 5（d））。有关详细信息，读者可以参考[2]，[8]，[21]。

3.2 特征提取

毫无疑问，选择好特征来识别手势路径在系统性能起着重要的作用。手势路径有三个基本特征：位置、方向和速度。我们分析这些从手势轨迹提取的特征的有效性，同时将它们结合起来测试它们的识别速率。手势路径是一个由手质心（xhand，yhand）组成的时空坐标。

在笛卡尔空间的该坐标可以直接从手势帧中提取。我们考虑两种位置特征。第一种位置特征是从质心到手势路径的各个点的距离 Lc ，因为同一手势根据不同的起点形成不同位置特征（公式 1）。第二种特征是计算从起点到手势路径上当前点的

距离 Lsc（公式 3）。

$$Lc_t = \sqrt{(x_{t+1} - C_x)^2 + (y_{t+1} - C_y)^2} \quad (1)$$

$$(C_x, C_y) = \frac{1}{n}(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n} y_t) \quad (2)$$

$$Lsc_t = \sqrt{(x_{t+1} - x_1)^2 + (y_{t+1} - y_1)^2} \quad (3)$$

其中 t=1,2,…,T-1，T 是手势路径的长度。(C_x, C_y)指在 n 点的重心。为了验证实时性，我们计算每帧后的手势路径的重心点。

第二个基本特征是方向，它给出了在手势处理过程中贯穿空间的手势走向。如上所述，方向特征是基于每个点上的手势位移向量的计算，它代表着手势路径质心的方向(θ_{1t})、连续两个点的方向(θ_{2t})和起点与当前手势点的方向(θ_{3t})。

$$\theta_{1t} = \arctan\left(\frac{y_{t+1} - C_y}{x_{t+1} - C_x}\right) \quad (4)$$

$$\theta_{2t} = \arctan\left(\frac{y_{t+1} - y_t}{x_{t+1} - x_t}\right) \quad (5)$$

$$\theta_{3t} = \arctan\left(\frac{y_{t+1} - y_1}{x_{t+1} - x_1}\right) \quad (6)$$

第三个基本特征是速度，它在手势识别阶段特别是在一些关键情形中起很重要的作用。速度是基于一个现实，在这现实里手势路径的拐点处手的速度降低。速度是通过欧氏距离除以视频帧里面的两个连续点的时间如下：

$$V_t = \sqrt{(x_{t+1} - x_t)^2 + (y_{t+1} - y_t)^2} \quad (7)$$

在直角坐标系中，我们使用不同的特征组合，以获得各种特征向量。例如，在帧 t+1 的特征向量可以通过位置特征(Lc_t, Lsc_t)、位置特征和方向特征(Lc_t, Lsc_t, V_t)、方向特征$(\theta_{1t}, \theta_{2t}, \theta_{3t})$、方向特征和速度特征$(\theta_{1t}, \theta_{2t}, \theta_{3t}, V_t)$、位置特征和速度特征与方向特征$(Lc_t, Lsc_t, \theta_{1t}, \theta_{2t}, \theta_{3t}, V_t)$来获得。

在时刻 t 每帧都包含一个特征向量集，其中的空间维数正比于特征向量的大小。在这种方式下，手势被描绘成一个有序的向量特征序列，它在三维中被处理和聚簇成一个离散值作为隐马尔可夫模型的输入。这可以通过使用 K-均值聚类算法[19]，[20]完成，它将手势模型在特征空间中分为 k 集群。

矢量量化：量化提取的特征得到了离散值。当诸如位置和速度等基本特征单独使用时，这些特征被规范化和乘上从 10 到 30 的不同标量值。另一方面，方向特征的规范值除以 10，20，30 和 40 已获得它的码字。除了直角坐标系中的组合特征，我们使用 K -均值聚类算法将手势特征分类到特征空间中的 k 集群。该算法是基于每个集群的中心到特征点的最小距离。我们将特征向量集分成集群集。这使我们能过通过一个集群在特征空间模拟手势轨迹。这计算得来的集群指数用来当做隐马尔可夫模型的输入。此外，在数据集中我们通常不知道集群的个数最好是多少个。为了得到在每个 K-均值算法中的每个执行中集群 K 的个数，我们假定 K=28，29，…，37，这样的假定是基于在所有的字母（A-Z）和数字（0-9）的分段部分的数量，其中每个直线段被编入同一集群。

假设我们有 n 个来自同一类得训练特征向量 x1,x2, ..., xn 的样本，同时我们将它们分为 k 集群，k<n。在集群 i 中我们让m_i为向量的平均值。如果这些集群被很好的分离，那么一个最小距离分离器被用来分离它们。也就是，如果$\|x-m_i\|$是所有 k 距离中的最小一个是，我们可以认为 x 是在集群 i 中。

•对均值 m1,m2, ...,mk 建立一个随机初始向量量化编码本

•在任何情况下都没有改变

我们使用估计方法来给每个训练特征样本分类到一个集群m_i中

For i=1 to k

对集群 i 用所有训练好的特征样本的均值代替m_i

end(for)

•end(until)

一个总的看法是，不同的手势有不同集群中的空间运动轨迹，而同样的姿态有非常相似的轨迹。

3.3 分类

在我们系统的最后阶段就是分类。在本阶段，Baum - Welch 算法[15]是用来对初始化隐马尔可夫模型参数做一个完整的训练来构建手势库。手势库中的字母（A-Z）和数字(0-9)的每个参考模型是通过左-右 Banded 模型根据它的复杂性用从 3 到 6 个不同状态来模拟。正如，如果训练样本数量跟模型参数相比不足的话，过

多的状态会形成过拟合问题。通过选择最大观察手势模型可能性分类手势路径。最可能的手势模型是所有 36 手势中观察可能性最大的手势。所观察到得手势 O 一帧一帧通过 Viterbi 算法识别（即累计，直到它收到手势信号结束）。下列步骤展示了 Viterbi 算法如何在手势模型 $\lambda_g(a^g, b^g, \Pi^g)$ 中工作的。（图 4）

1. 初始：

$$for\ 1 \le i \le N,$$

$$\delta_1^g(i) = \Pi_i^g . b_i^g(o_1)$$

2. 递回（累计观察概率计算）：

$$for\ 2 \le t \le T,\ 1 \le j \le N.$$

$$\delta_t^g(j) = \max_i[\delta_{t-1}^g(i).a_{ij}^g].b_j^g(o_t)$$

3.终止：

$$P(O|\lambda_g) = \max_i[\delta_T^g(i)]$$

其中，N 是状态个数，Π_i^g 是状态 i 的初始值，a_{ij}^g 是状态 i 变成状态 j 的权值，$b_j^g(o_t)$ 指在在 t 时刻在状态 j 下 o 发散的权值，而 $\delta_t^g(j)$ 指在 t 时刻状态 j 下最大可能值。

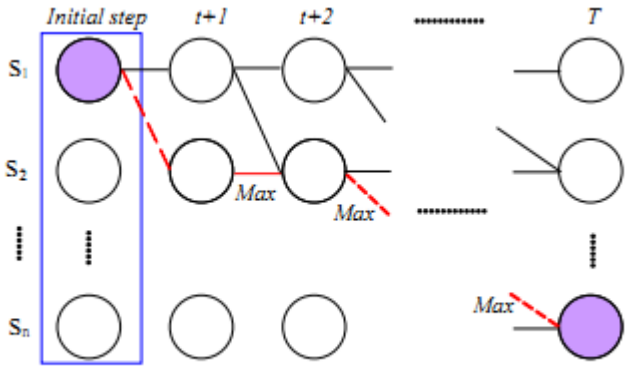


Fig. 4. The best path for LRB model with N states where it starts from S_1 to S_N, N=3, 4, ..., 6 and t=1.

4. 实验结果

我们系统已经用不同的视频序列包括重叠和部分遮挡视频来测试过。我们提出的系统通过隐马尔可夫模型使用单只手的运动轨迹能够实施执行和展示良好识别三维彩色图片里的字母和数字结果的能力。输入图像通过拥有有 6 毫米 15fps 的 240×320 像素的图像分辨率的大黄蜂立体摄像机系统捕获，Matlab 实现。在我们的实验结果，每一个孤立的手势是基于 30 的视频序列，其中 20 个样本进行训练和 10 个样本进行测试（总共，我们的数据库包含 720 个训练样本和 360 个测试样本）。手势识别模块将手势路径跟参考手势库进行匹配，以识别出手势属于哪类。通过 Viterbi 算法采用 LRB 拓扑从 3 到 6 不同数量状态一帧一帧地实时识别字母和数字以计算更好的优先级。

我们测试了直角坐标系中三个基本特征（位置、方向、速度）的重要性。此外，对隐马尔可夫模型观测序列或在单独特征下正常量化或在组合特征下采用 k-均值聚类算法量化。从表 1，孤立的手势识别率在采用特征$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$下获得了最佳效果 98.33%。识别率是正确识别手势数目与测试手势的数量（公式 8）。

$$Reco.\ ratio = \frac{\#\ recognized\ gestures}{\#\ test\ gestures} \times 100\% \qquad (8)$$

TABLE I
RESULTS OF HAND GESTURES ACCORDING TO DIFFERENT FEATURES EXTRACTION IN CARTESIAN SYSTEM WITH THE BEST FEATURE CODE NUMBER.

Feature type	Feature space	# Feature code	Training data	Hand gestures results		
				Testing data	Correct data	Recognition (%)
Separated in Cartesian coordinates	Lc	20	720	360	187	51.94
	Lsc	25	720	360	118	32.78
	v	25	720	360	206	57.22
	θ_1	18 ; 36	720	360	349	96.94
	θ_2	18	720	360	321	89.17
	θ_3	18	720	360	236	82.22
Union in Cartesian coordinates	(Lc, Lsc)	35	720	360	293	81.39
	(Lc, Lsc, V)	33	720	360	309	85.83
	$(\theta_1, \theta_2, \theta_3)$	33	720	360	338	93.89
	$(\theta_1, \theta_2, \theta_3, V)$	34	720	360	339	94.16
	$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$	33	720	360	354	98.33

表 1

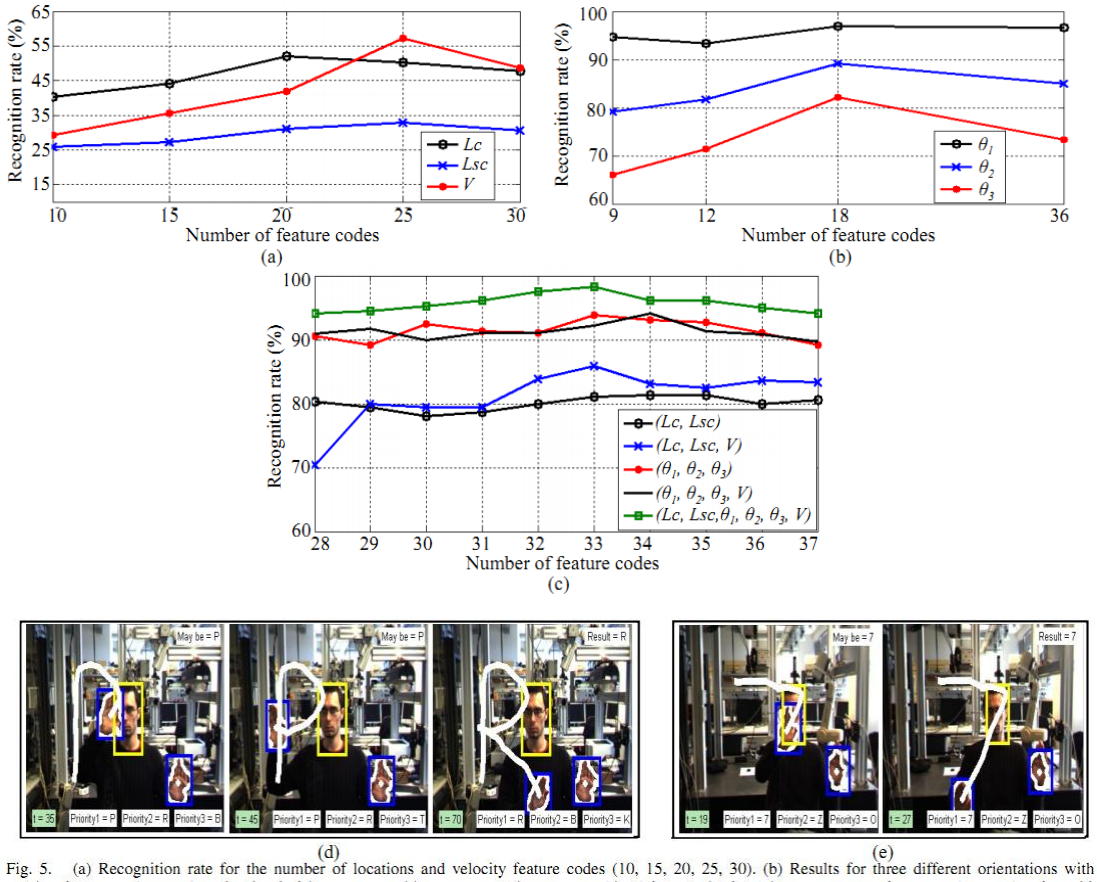


Fig. 5. (a) Recognition rate for the number of locations and velocity feature codes (10, 15, 20, 25, 30). (b) Results for three different orientations with varying feature codes number (9, 12, 18, 36). (c) Recognition rate according to a combined features in Cartesian system over feature codes number from 28 to 37. (d) The high priority is alphabet 'P' at t=45 and at t=70 the result is 'R'. (e) Solving overlap problem between hand and face at t=19 and the high priority is '7' at t=27.

图 5

根据图 5(a)和 5(b)中单独特征，方向特征$(\theta_1, \theta_2, \theta_3)$比位置特征$(Lc, Lsc)$或速度特征(V)的识别率要好。这导致方向特征$(\theta_1 = 96.94\%)$是三个基本特征中最有效的一个。此外，速度特征展示比方向特征更低的识别力（57.22%）。同时，位置特征Lsc有最低的识别率 32.78%。此外，包括速度信息的特征(Lc, Lsc, V), $(\theta_1, \theta_2, \theta_3, V)$和$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$比使用速度特征有更高的识别率（图 5(c)）。简而言之，图 5 说明了实验的结果表明最好的特征码号（对特征$(Lc, Lsc, \theta_1, \theta_2, \theta_3, V)$来说，最好的特征码号是 33）。相应的，图 5(d)和图 5(e)展示了手势字母'R'和数字'7'的系统输出，同时用 3D 深度图解决了手和脸的重叠问题。

5. 结论

在本论文中，我们提出了一个通过使用隐马尔可夫模型（HMM 模型）能够实时从彩色图像的序列中识别字母字符（A-Z）和数字（0-9）的系统。这个系统使用笛卡尔坐标系统中的位置、方向和速度的组合特征。我们已经表明，这些特性的有效性可以产生合理的识别率。数据库包含孤立手势的 720 帧训练样本和 360 帧测试样本。当应用于几个包括诸如部分遮挡和重叠的复杂情形视频样本时，这个系统能表现出良好的性能。结果表明，本系统是适合实时应用，并且具有 98.33％的手势识别率。未来研究将采用指尖与多摄像机系统结合来识别手势点和句子而不是手势重心点。

鸣谢

这项工作由 Bernstein-Group BMBF(FKZ: 01GQ0702), 文理学院助学金(C4-NIMITEK2,FKZ:UC4-3704M),Forschungspraemie(BMBF-Frderung,FKZ:03FPB00213)和 DFG 建立的合作研究中心”Companion-Technology SFB / TRR 62”支持。

摘 要

研究人、计算机以及它们之间相互影响的技术称之为人机交互，是人与计算机通过人机界面进行某种形式上的信息的交流以完成一定的交互任务的过程。人机交互技术已经从以计算机为中心逐渐的转变为一人为中心，手势的交互是一种很容易学习的、自然、直观的人机交互手段，手势交互界面有着广阔的应用前景。近年来，以手机、PDA和掌上电脑为代表的手持移动设备得到了日益广泛的应用，手持移动计算机已经逐渐成为当今世界的主流计算模式之一。随着移动设备自身的软硬件性能的提高和带宽、网络覆盖等条件的改善，人机交互的效率过低和自然性不好的问题暴露得越来越明显，计算机应用的主要障碍也已由硬件技术转变为人机交互和用户界面，基于视觉的手势识别技术已经成为一个研究的重点。

本文总结和介绍了现有的手势识别技术，手势识别研究的关键内容以及手势识别技术的发展历史。接着，本文主要对手势识别涉及的主要技术进行了研究。

关键词：手势；识别；跟踪；人机；交互

摘　要

1 绪 论

1.1 课题背景

计算机系统是由人、计算机软件、计算机硬件来共同构成的人机系统。人机交互研究的是人和计算机之间相互影响的技术，是人与计算机通过人机界面进行某种形式上的交流用来完成一定交互任务的一个过程。人与硬件、软件的交叉部分就构成了人机界面。人机界面是介于用户和计算机系统之间，是计算机与人之间传递、交换信息的媒介，是人们使用计算机系统的综合操作环境。它作为计算机系统的一个重要的组成部分，是计算机科学、认知科学、心理学的交叉研究领域，也使计算机行业竞争的焦点从硬件转移到软件之后，又一个新的、重要的研究领域。随着计算机系统的发展，用户界面的发展经历了批处理、联机终端、菜单等阶段，现在正处于以图形用户界面为主流的阶段。交互式系统的发展趋势正逐渐以“以机器为中心”转移到“以人为中心”、“人际和谐交互”的方向上。而“以人为中心”的人机交互的一个重要研究方向，就是通过模拟与人类类似的感知类型进行信息传递，这些研究包括人脸识别、面部表情识别、头部运动跟踪、手势识别、以及体势识别等等[1][2]。

1.2 研究意义

手势语言是一种靠视觉和动作进行交流的一种特殊的语言，它还是一种包含信息量最多的人体语言，它与语音及书面语等自然语言的表达能力相同，因而在人机交互方面，手势完全可以作为一种有效的、自然、直接的交互手段，具有很强的表意能力，可以在很多特殊的场合表达一些特定的信息。而基于计算机视觉的手势识别技术应用于人机交互接口具有用户友好、直接而有效等等优点。这使得手势交互可以成为人机交互过程中的一个非常自然的、直观的交互通道，符合“人际和谐交互”及“以人为中心”的人机交互发展方向。

作为一种自然，直接的交互方式，基于视觉的手势交互方式有着广泛的应用前景，现有的主要领域包括：

1. 在控制机器人和机器人远程操作中的应用。例如：在伊拉克战争中使用的

智能机器人去拆炸弹，远程医疗的远程控制领域。

2. 辅助聋哑人生活。辅助聋哑人的生活，通过手语界面可以减少聋哑人在生活中的障碍。
3. 在电子游戏领域。现在更多的游戏实现了手势识别来代替以前的按键控制，实现了人机交互的综合应用。例如：游戏厅里面的跳舞系统，就是人机结合的良好平台。

1.3 论文主要的研究内容

本文主要研究了基于视觉的手势识别技术，重点研究的是手势的分割、手势的跟踪、手势特征提取和手势识别算法。

首先，本文选择 HSV 空间的 H 分量等信息对手势的原始图像进行分割，并对图像进行平滑滤波和形态学处理，得到完整的手势二值图像，通过八领域搜索法计算手势的轮廓。之后，采用 Camshift 算法对手势进行跟踪。最后，根据手势图像和轮廓的结构信息和统计信息，对手是进行识别。

2 基于视觉分析的手势图像预处理

2.1 手势识别概述

2.1.1 手势识别的定义和分类

通常我们把手势的定义分为:手势是手或者手和臂结合产生的各种姿势和动作，以助于表达情绪、想法或强调所说的话。根据不同的标准，手势还有着不同的分类：

根据手势的空间特性，手势可分为动态手势和静态手势。动态手势强调的是手在做一个动作的一个过程，表现为手在一个时间段上的手部动作的姿势的一个序列；静态手势的意思是在某一个时刻点上手在一定空间的姿势，包括朝向、手形、与身体的相对位置。

首先，对于静态手势，是通过八连通搜索法来计算手势的轮廓，并根据手势轮廓的特点，研究了手势特征提取的方法，提出了统计特征、结构特征结合的特征提取方法。对于动态手势，本文讨论了手势跟踪技术，采取了 Camshift 算法对手势来进行跟踪。在手势识别方面利用不同特征的差异，简化了识别计算，来提高系统的实时性，所以本文提出了一种分层识别的方法，这是一种根据手势特征值特性来设计的识别方法。在动态识别方面，按照自然交互的要求，设计了简单易用的分区域识别法。

所以将手势按功能做如图 2.1 所示的划分：

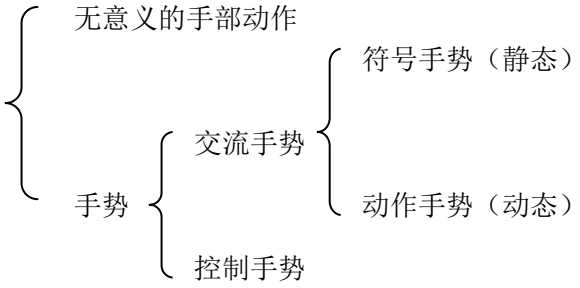


图 2.1　手势的分类

在人机交互领域中，根据手势的所表达含义可以将手势分为两类：一种是无意义的动作，一种是传递这用户意图的手势。根据功能的特点，手势又可分为控制手势和交流手势。交流手势的本质就是为了更好的传递信息，比如交谈中的伴随着不同的手势，包括具有语言描述作用的一些符号手势以及表示指示方向的手势等等。控制手势是用来控制环境中的物体，如平移，旋转，托起等等。

对于手势识别系统的分类，根据手势的采集设备可以把手势识别系统分为两类：基于视觉的手势识别系统和基于数据手套的手势识别系统。

基于数据手套的手势识别系统是最早的手势识别系统，由于当时机器视觉技术以及视觉采集设备的限制，需要用户佩戴数据手套，通过数据手套来测量出手指或者手臂的关节角度和位置等信息，进而来识别用户的手势。采用数据手套手势识别系统的优点是采集装备的应用使手势建模的难度大大降低，同时采用数据手套对手势信息的采集的有效性较高，使得基于数据手套的手势识别系统有实时性和准确性的特点。然而这种方法对于用户而言，在人机交互的过程中必须佩戴昂贵而且比较笨重的手势采集装备，是一种间接的交互方式，限制了手势交互的自由性，这就导致了基于数据手套的手势交互方法无法成为一种自然的交互方式。

随着目前计算机视觉技术的发展，基于视觉的手势识别技术也越来越成熟，它主要通过摄像机来采集手势的视觉信息，从视频图像中提取手势，并进行识别，用户不需要佩戴任何的设备，可以直接与计算机之间进行交互，与基于数据手套的手势识别技术相比，从视觉信息中要完整地恢复出原始的手势信息难度相对较大，可识别的识别率和手势数量、实时性方面还不能达到基于数据手套手势识别的效果。因为基于视觉的手势识别技术对输入设备的成本低，对用户的限制少，人手是处于一种自然状态，使人能够以自然的方式与计算机之间进行交互的优点，所以基于视觉的手势识别技术符合人机交互技术发展的方向，也是未来手势识别技术发展的趋势和目标。因此本文主要关注的是基于视觉的手势识别技术。

2.1.2 基于视觉的手势识别技术的主要研究内容

手势建模、手势分析、手势识别是基于视觉的手势识别技术的三个核心部分[3]。

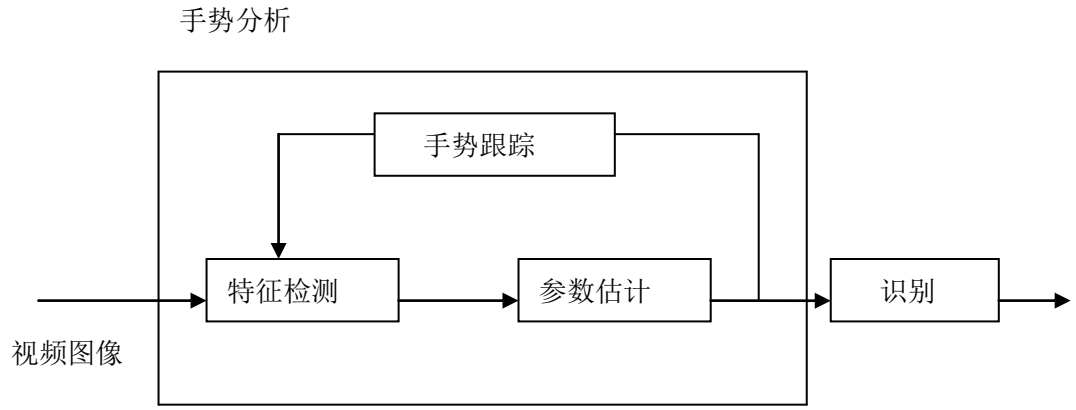


图 2.2 手势识别系统流程图

在建立了手势建模的情况下，手势识别系统的流程图如 2.2 图中所示。手势建模的意思是选取什么样的模型来描述所要表达手势，这个数学模型包含了手势的时间特征和空间属性。一旦模型确定好了之后，接下来的任务就是从单幅或序列图像中计算出模型的参数，这些参数描述了手的姿势和运算的轨迹。手的定位、跟踪以及选择合适的图像特征是手势分析阶段的关键。手势识别阶段就是将手势特征参数与已知的模型进行一一匹配，从而来判断手势的类别和属性。

1.手势建模

手势建模对于手势识别系统至关重要的，特别是对确定识别范围起关键性作用，一般来说手势建模方法被分为两大类[4]：基于表观的手势建模和基于 3D 模型的手势建模。前者是直接从观察到的视频图像去推断手势；而后者考虑了手势产生的中间媒介（手和臂）。图 2.3 是对两种建模方法的进一步分类。

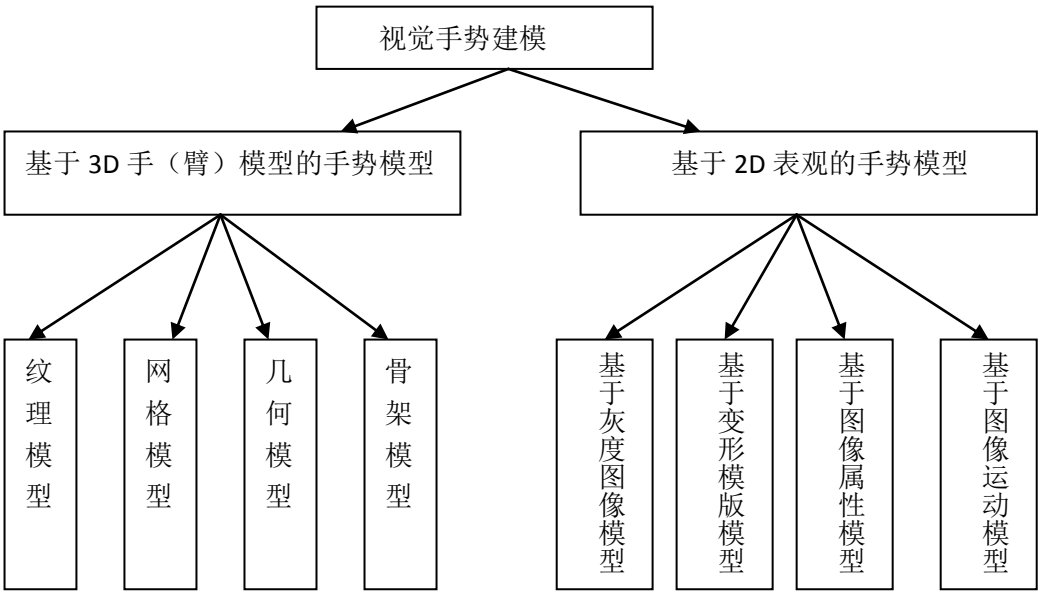


图 2.3　手势模型的分类

2.手势分析

在模型确定之后，手势分析的任务是估算所选手势模型的参数，由手势跟踪、收拾图像分割、参数和特征检测估计几个部分组成。

在进行特征检测的过程中，首先要从场景中分割出手势。手势图像分割技术有基于运动信息、基于颜色信息和基于综合信息这三种方式。基于运动信息的分割方法是利用图像差分法把用户的手部区域从背景中分离出来，常用的差分法有帧间差分法和背景差分法。基于颜色信息的手势图像分割技术利用肤色在颜色空间中的分布特性，利用阀值分割的方法将手势分割出来。第三种方法则综合利用肤色信息和运动信息来定位用户的手势。

在手势图像分割结束之后，为了提取手势的动态信息，同时避免在视频序列中每一帧的定位手势以节省计算的资源，需要在场景中跟踪用户的手势。在动态手势识别中，手部动作速度比较快，一般在 5m/s 以上，为了达到人机交互界面所需要的准确性和实时性的要求，需要一种在手部变形和复杂背景下跟踪实时性好而且鲁棒性高的图像跟踪算法。目标检测算法和运动估计算法是视频目标跟踪算法主要的两个部分。在目标检测算法方面，常用的方法有光流法、背景分差法等。运动估计算法主要包括粒子滤波算法、卡尔曼滤波算法以及在此基础上的各种改进的算法。

在结束特征检测之后，根据所选用的手势模型来估计模型参数。不同的手势

模型需要估计不同的模型参数，但是用于计算模型参数的图像特征通常都是基本相似的。常用的图像特征包括二值图像[6]、灰度图像[7]、边界[8]、区域[9]及轮廓[10]或者指尖[11]等。对于人机交互的手势识别系统，提取的手势特征首先要能够区分不同手势，另外还需要对手势的旋转、平移、缩放等变化能够很快的适应。

2.2 基于肤色的手势图像分割

手势图像分割的第一步是基于视觉的手势识别过程，是最为重要的一步，手势图像分割的好坏直接影响到后面的手势跟踪、手势特征提取以及手势识别的结果。手势图像分割就是将有意义的区域—手势从摄取的手势图像中划分出来，可以通过以下几种方法来实现：

1. 增加限制

通过简化背景，比如使用白色或者黑色的墙壁、深颜色的服装作为手势图像采集的背景。或者戴特殊的手套，通过强调前景来简化手和背景域之间的划分，加深两者之间的对比。但是这些人为附加的限制影响了手势交互的失去了自由性。

2. 模型的比较

首先建立手势在各个时刻、不同比例、不同位置下的手形图像，然后通过匹配的方法来实现手势的分割。这种方法的缺点是计算量非常大，而且无法实现实时识别。

3. 轮廓的跟踪

典型的有基于 Snake 模型的手势分割，利用 Snake 模型对噪声和对比度的敏感性来有效跟踪目标的形变和非刚体的复杂运动，达到将目标从复杂背景中分割出来的目的，这种方法的效果比较好，但同样无法用于实时系统。

4. 目标背景相减法及其改进算法

就是将目标图像和背景图像相减，此方法对消除背景图像具有很明显的效果，但要求已知背景并且背景不变，这一点限制了算法的适用范围。

5. 基于肤色的分割

主要根据肤色在颜色空间中分布的特点，通过快速的找到手可能运动的候选区域，缩小后续检测的范围。从背景图像中分割出肤色区域，用肤色特征信息来实现手势和背景的分离。基于肤色的分割方法有着直观、高效并且准确的有点，

也是本文采中手势图像分割所采用的方法。

由于颜色是人手表面最为显著的特征之一，所以在计算机视觉技术中利用颜色检测人手是一个很自然的想法，然而在不同环境下，不同的用户的肤色表现有着很大的差别，所以如果想要实现良好的肤色分割效果，那么首先需要选择一个肤色可以良好聚类并且可以适应光线环境变化的颜色空间。

手势图像预处理模块主要包括了对图像的平滑去噪、基于 H 分量的手势分割及对手势二值图像的形态滤波以及手势轮廓的提取。预处理的目的是提取手势的二值图像和轮廓，为手势特征提取提供条件。预处理的流程为图 2.4 所示：

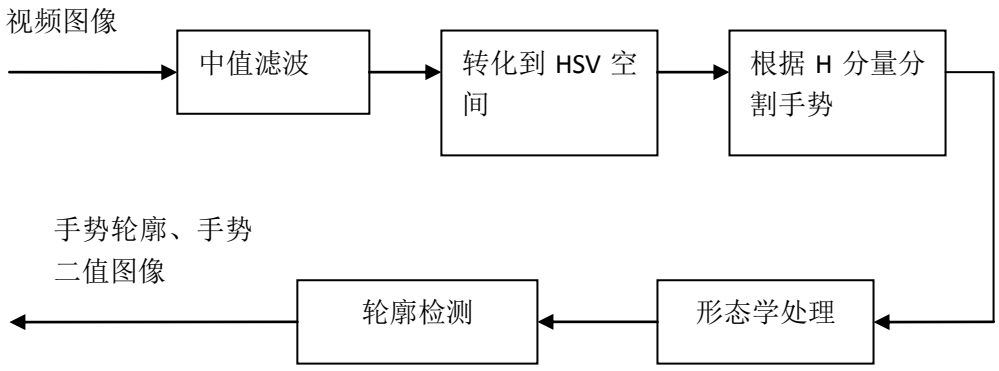


图 2.4 手势预处理模块流程图

2.2.1 肤色检测所采用的颜色空间

颜色空间指的是颜色的数学表示方法，用来指定和产生颜色，让颜色更加的形象化。颜色通常是用三维模型来表示，用三维坐标来指定空间中的颜色，在一些特定的颜色空间中，一个指定的三维坐标对应着颜色空间中的一个特定的颜色，常用的颜色空间有：RGB 颜色空间、YXZ 颜色空间、YUV 颜色空间、HSV 系列颜色空间。

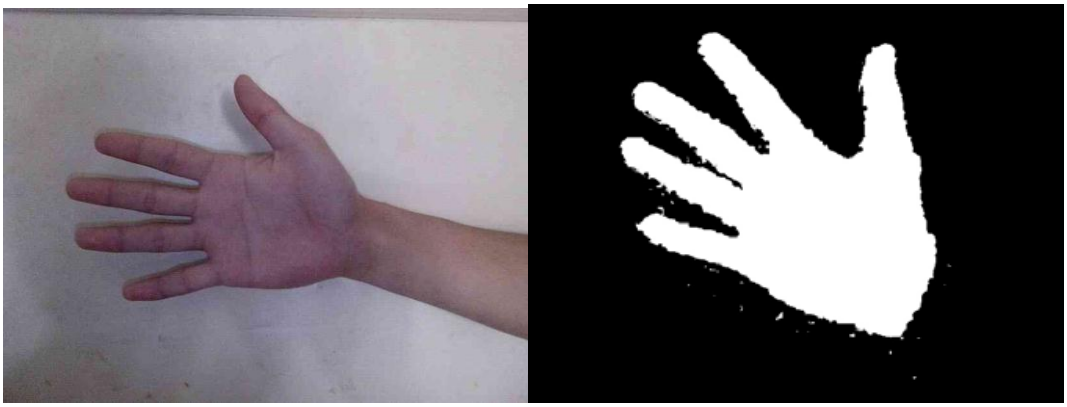
由于 HSV 空间中的 H 和 S 分量是独立于 V 分量的，也就是说颜色和饱和度信息与亮度信息是互相独立的，因此可以有效的减小受到光线的影响，所以在本文中采取 HSV 系列颜色空间来对手势进行分割，使得对肤色的分割算法可以通过更简单的方法来进行，这样就有利于增强系统的实时性。

2.2.2 HSV 空间下的手势图像分割

本文采用的是 HSV 空间中的 H 分量来进行肤色的统计。具体的步骤是对 100 幅不同光照的情况下不同的人手部的肤色图像进行统计，第一步：将 RGB 格式的图像转换成为 HSV 的格式。第二部：统计每个像素的 H 值在颜色空间分量中的取值范围内的分布范围直方图。根据对手部的肤色信息，可以对采集到的包含手部的原始的图像进行二值化处理，从而得到人手的区域。根据原图想（x，y）处的 H 值和 R、G、B 值判断二值图（x,y）处的像素值为：

$$f(x,y)=\begin{cases}255 & \mathrm{H}\in\{2,20\}\ \text{and R>G>B}\\ 0 & \text{else}\end{cases} \qquad (2,1)$$

式 2.1 中 $f(x,y)$ 是二值化图像中坐标为 (x,y) 的像素值，H、R、G、B 分别为原图像中 (x,y) 处的像素的 RGB、HSV 颜色空间所对应的分量的值。经过处理之后，就可以得到手势区域的二值图，其结果如 2.1 图所示。



(1) 简单背景下的手势图像分割

(2) 复杂背景下的手势图像分割

图 2.1 手势原图及二值化后的图像

图 2.1 中(1)为简单背景为单一颜色下的手势图像的分割，(2)为复杂背景下基于肤色的手势图像的分割。从实验的结果可以看出，基于 H 分量的肤色分割在简单背景和复杂背景下都是可以实现稳定的手势图像的分割，从而得到完整的手势二值化图像，但是图像的分割会受到视频采集设备的性能和光线等的影响，当有噪声存在的情况下，分割出来的手势图像边缘比较粗而图像内部存在空洞，解决这个问题需要后续再对图像进行滤波处理。

2.3 图像平滑

在图像的采集、传输过程中，无法避免的都会受到一些外界环境的影响，例如：噪声。那么得到的图像画质会因噪声而在不同程度上出现一些变异，因此我们必须首先对图像进行平滑操作，过滤掉部分噪声的影响。

图像的平滑是一种最常用的数字图像处理技术，主要是为了减少噪声对图像的影响。一般情况下，在空间域内可以使用领域平均来减少噪声：在频率域，由于噪声频谱通常都分布在高频段，因此可以采用各种形式的低通滤波的办法来减少噪声。常用的图像平滑方法有：中值滤波、领域平均、频域平滑技术。

根据系统视频采集设备的特点，为了需要解决采集到的原始图像中的噪声对手势图像二值化的影响，所以本文采用中值滤波技术进行平滑的处理，在消除噪声影响的同时还保留了原图想的细节。

(1)原始图像

(2) 未经平滑处理的二值化图像

(3)经过中值滤波处理后的二值化图像

图 2.2 图像中值滤波处理后的实验结果

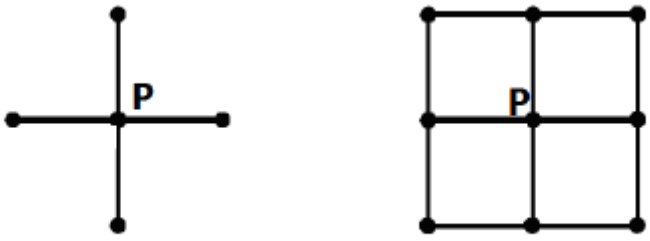
图 2.2 是本文中所采用的中值滤波对直接分割出来的二值化的图像进行处理后的结果，图中(1)为原始的图像，（2）是根据上文分割方法所得到的手势图像，(3)是用 5x5 模版对二值图像进行中值滤波处理后的结果。从上面的图像我们可以看出，在直接分割得到的二值图像中，是存在严重的噪声干扰的；经过中值滤波处理之后，就去掉了大部分的噪声的影响。不过在处理之后的图像中，虽然去除了绝大部分的噪声，手势图像的边缘还是比较粗糙的，并且还有是一些空洞的存在，为了解决这个问题，可以通过形态学滤波对图像进行进一步的处理。

2.4 形态学滤波

在处理图像二值化的过程中，本文采用形态学方法对得到的二值化图像进行滤波。因为判断肤色的标准是按照经验得到的统计信息，并不能完全准确的判断出每一个像素的性质：手部的边缘部位易受光照的影响，同时也容易受到噪声的影响，H 值不稳定，所以二值化得到的图像无法避免噪声、孔洞和粗糙的边缘的存在。

在数字图像处理的过程中，由于数学形态学算法有填充洞孔、平滑轮廓、连接断裂区域等特性，常常被用在处理各种图像操作中。数学形态学是一种以形态为基础从而对图像进行分析的一种数学工具。其基本运算分为 4 种，即膨胀、腐蚀、开运算和闭运算。通常都是用集合论名词来定义的，并用专门定义的结构元素对图像进行一定的操作。图 2.3 中为应用最为广泛的由 4 连通的 3x3 领域（5 点）和 8 连通的 3x3 领域（9 点）组成的结构元素。基于这些基本运算还可以推导和组

合各种数学形态学实用算法。



2.3 常用的结构元素

对于本实验中的二值化手势图像，存在边缘毛刺和内部的空洞，我们可以结合开运算和闭运算去除这些空洞，如图 2.4 所示的是对一幅包含噪声、内部小空洞和边缘毛刺的二值化之后的手势图像进行开闭运算的结果。其中(1)是对肤色分割得到的手势二值图像在经过中值滤波处理之后得到的图像，（2）是对(1)中图像用 5x5 结构元素进行形态学处理后的结果。

(1) 经过中值滤波得到的手势的二值图像

(2) 经过开、闭运算得到的手势二值图像

图 2.4 二值手势图像开闭运算后结果

从上图中可以看出，中值滤波处理之后的图像中还是包含了少量的孔洞和噪声，而孔洞和噪声的尺寸和手势相比是比较小的，经过开闭运算处理之后，孔洞和噪声得到了有效的减少。

2.5 边缘检测与轮廓提取

在得到了手势的二值图像之后，为了简化图像的信息，突出手势的结构特征，要对图像进行轮廓提取和边缘检测。这两种图像处理方法都是从图像中提取目标物体的形状，用提取的形状来描述特定的手势。其目的是把图像中人们感兴趣的部分分离出来，突出想要的最终目标，减少处理的信息量。两者区别在于边缘检测是一种并行的检测方法，提取图像中所有目标物体的边缘，同时检测出来的边缘有可能不是封闭的，而对于轮廓提取是一种串行的方法，提取图像中目标物体最外层的轮廓，提取出的轮廓是封闭的曲线图形。

2.5.1 边缘检测

图像的基本特征就是图像的边缘，边缘指的就是它周围像素灰度有阶变化或者屋顶变化的像素的一个集合。

Canny 边缘检测器是最为精确的一种边缘算子，在现在已经得到了广泛的应用，它按照优良检测、精确定位和对边缘单一响应这三个标准，Canny 边缘检测器

是通用性最优的一种方法。Canny 检测主要检测的是阶跃性边缘。它的基本思想就是在图像中找出具有局部最大梯度幅值的那个像素点，检测边缘的主要工作是寻找能够用于实际图像的梯度数字逼近。由于实际的图像是经过了摄像机的光学系统和电路系统固有的低通滤波器的平滑，所以，图像中的阶跃性边缘不是十分的独立。图像也容易受到噪声和场景中不希望出现的一些其他因素的干扰。

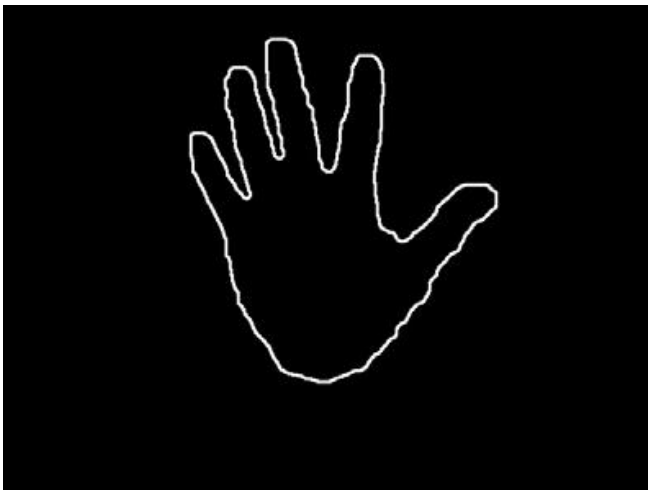
2.5.2 轮廓提取

边缘检测是一种并行的处理技术，检测出来的边缘往往不是封闭的，而轮廓提取它是一种串行的检测技术。其基本的方法是：先根据一些严格的“探测标准”找出其目标物体轮廓上的像素点，再凭这些像素点的某一些特征用一定的“跟踪标准”找出目标物体其他的像素点。

在试验中，为了方便对手势特征的提取，就采用了八领域搜索法对手势二值图像的轮廓进行提取，得到以链码形式表示的手势图像的轮廓。

手势轮廓提取效果如图 2.5 所示

(1) 手势二值图像



(2)八领域搜索法提取的手势轮廓

图 2.5 手势轮廓提取效果图

图 2.5 中，(1)是通过上面的方法所提取得到的手势二值图像，(2)图是用八领域搜索法根据手势二值图像提取的手势轮廓，并且根据此轮廓信息绘制出的手势轮廓图像。从结果中我们可以看出，通过八领域搜索法从手势二值图像中准确的提取出手势的轮廓。

2.6 本章小结

本章详细介绍了手势图像的预处理的问题，包括了对肤色空间的选择，对分割得到的二值图像的形态学滤波、手势图像的分割、对原始图像的平滑以及手势图像轮廓的提取。手势图像的预处理的目的是为了对下一步将要进行的手势分析提供一个良好的手势模型，对于手势预处理质量的好坏直接关系到手势分析的准确度，对整个识别系统的效果也是至关重要，是所有后续工作的基础。