ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

Hand Gesture Spotting and Recognition in Stereo Color Image Sequences Based on Generative Models

* Fayed F. M. Ghaleb, **E. A. Youness, ** M. Elmezain and ** F. Sh. Dewdar
*Mathematics Department, Faculty of Science, Ain Shams University, Cairo, Egypt
* * Computer Science Division, Faculty of Science, Tanta University, Tanta, Egypt

Abstract*—This paper proposes an automatic system that handles hand gesture spotting and recognition simultaneously with no time delay based on a generative model as Hidden Markov Models (HMMs). Color and depth information are used to detect hands and face. The depth information computed from stereo camera system is used to identify the region of interest, neutralize completely complex background, as well as illumination variation and it also increases the accuracy of objects segmentation. To spot meaningful (key) gestures of numbers (0-9) accurately, a stochastic method for designing a non-gesture model with HMMs is proposed without training data. The non-gesture model provides a confidence measure that is used as an adaptive threshold to find the start and the end point of meaningful gestures, which are embedded in the input video stream. Experimental results show that, the proposed system can successfully spot and recognize meaningful gestures with 93.31% reliability for HMMs. In addition, the model inference by HMMs is faster and the saving time is 66.42% using relative entropy.

Index Terms— **Gesture Spotting, Gesture Recognition, Pattern Recognition, Computer Vision.**

I. INTRODUCTION

The process of communication is to transfer information from one entity to another. Naturally, hand gestures are powerful human-to-human communication channel, which forms a major part of information transfer in our everyday life. There are many ways to perform and interpret a human action using either hands and/or arms. A gesture is a spatio-temporal pattern which may be static, dynamic or both. Hand gestures are easy to use and more convenient for humans to interact with computers. For example, sign languages are considered as one of the main applications areas which have been used among the deaf people (i.e. speech-disabled people) [1]. In addition, the people with the ability to speak also use gestures in order to communicate with each other. There are many successful applications of hand gesture recognition like human-robot interaction [2], television control and computer game [3], video annotation and indexing [4], and video surveillance [5]. The task of extracting meaningful patterns from input signals is called pattern spotting [6], [7]. In gesture spotting, an instance of pattern spotting is required to locate the start point and the end point of a gesture (Fig. 1). The gesture spotting has two major challenges that arise in hand gesture recognition: segmentation [1], [8] and spatio-temporal variabilities [9], [10]. The segmentation problem is about determining the start and the end point of the gesture in a continuous hand trajectory. As the user switches from one gesture to another, his hand makes an intermediate move linking the two consecutive gestures. A gesture recognizer may attempt to recognize this inevitable intermediate motion as a meaningful one. The other difficulty of gesture spotting is caused because the same gesture varies dynamically in shape, trajectory and duration even for the same person. Therefore, the recognition step should consider both the spatial and temporal variabilities simultaneously. A robust recognition phase extracts the gesture segments from the input signal and matches them with the reference patterns regardless of the spatio-temporal variabilities. The latest advancements in computer vision and computer hardware technologies make the research of real-time hand tracking and gesture recognition promising. However, many current approaches still suffer from the limitation of accuracy, robustness and speed. This makes the gesture interaction indirect and unnatural. To face the mentioned challenges, a forward gesture spotting method is proposed, which simultaneously handles the hand gesture spotting and recognition in stereo color image sequences without time delay. To spot meaningful gestures accurately, stochastic method for designing a non-gesture model with HMMs are proposed with no training data. The non-gesture model provides the confidence measure and is used as an adaptive threshold to find the start and the end points of meaningful gestures which are embedded in the input video stream. One of the main contributions of this paper is to exploit depth image sequences.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

The main motivation behind the use of depth information is to identify the Region of Interest (ROI) without processing the whole image, which consequently reduces the cost of ROI searching and increases the processing speed. Furthermore, the depth information is used to resolve complex backgrounds (i.e. neutralize complex background) completely, as well as illumination variation and it also increases the accuracy of objects segmentation. In the case of overlapping (i.e. ambiguities) between the hands and face, the depth information is used to identify the objects under occlusion. Additionally, the non-gesture model with HMMs is modified by using relative entropy function to cure the problem of increasing number of states. The main objective is to save time and space, and to increase the spotting speed.

II. HAND GESTURE SPOTTING SYSTEM

We propose a real-time system that recognizes meaningful (key) gesture for numbers (0-9) in stereo color image sequences using generative models. The main processes of the proposed system are illustrated in Fig.1.

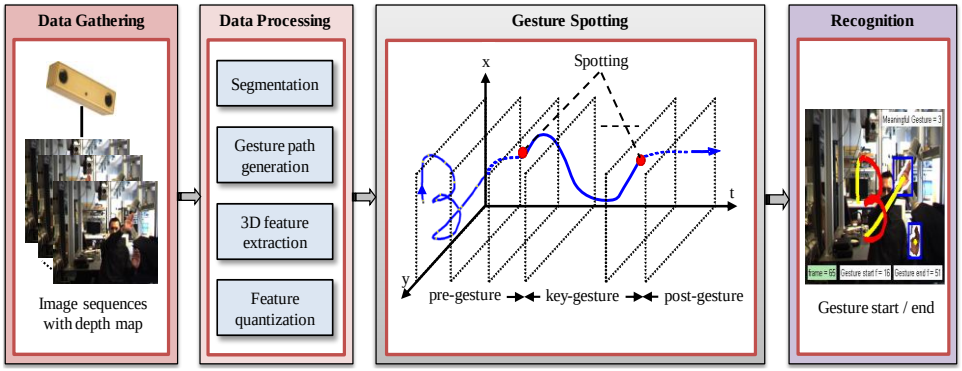

Fig. 1. Concept of the hand gesture spotting and recognition system

A. Preprocessing

Our main motivation is to improve the gesture recognition in natural conversations. This requires powerful techniques for skin segmentation and occlusion handling between hands and face to overcome the difficulties of overlapping regions. Therefore, a method for detection and segmentation of the hands in stereo color images with complex background is employed in which the hand segmentation and localization takes place using depth map and color information. This stage contains two steps: skin segmentation using Gaussian Mixture Models (GMMs) with *YCbCr* color space and hand localization using blob analysis like *region props function* [12], [13]. The following sections describe these parts.

1. Automatic Segmentation via GMMs

Segmentation of skin colored regions becomes robust if only the chrominance is used in analysis. Therefore, *YCbCr* color space is used in our system where *Y* channel represents brightness and (*Cb, Cr*) channels refer to chrominance. The channel *Y* is ignored in order to reduce the effect of brightness variation and only the chrominance channels are used which fully represent the color information. For more details, the reader can refer to [14], [15].

2. Depth Map

Image acquisition step contains 2D image sequences and depth image sequences. For the skin color segmentation of hands and face in stereo color image sequences, an algorithm is devised which calculates the depth value in addition to skin color information. The depth information is gathered by passive stereo measuring based on mean absolute difference and the known calibration data of the cameras. Several clusters are composed from the resulting 3D points. The clustering algorithm is considered as a kind of region growing in 3D which uses 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 [16], [17].

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

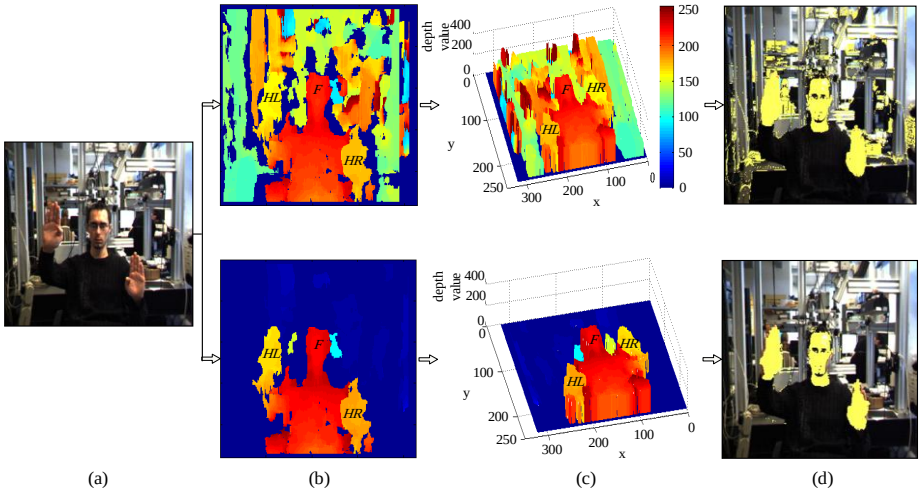

Fig. 2. (a) Original 2D image. (b) Normalized 2D depth image. (c) Normalized 3D depth. (d) The top image represents skin pixel detection with depth value up to 10 m. In addition, the skin pixel detection without noise is represented in the bottom image (the depth value ranges from 30 cm to 200 cm). Yellow color shows skin pixels detection. *F* refers to the face, *HL* and *HR* represent the left and right hands respectively.

The classification of the skin pixels is improved from the top images in Fig. 2 by exploiting the depth information which contains the depth value associated with 2D image pixel. The depth information is used to identify the region of interest without processing the whole image which consequently reduces the search cost of a region of interest and increases the processing speed. The depth value lies in the range from minimum depth 30 cm to maximum depth 200cm in our application. However, the depth range is adaptively varied according to ROI. By the given 3D depth map from camera set-up system, the overlapping problem between hands and face is resolved since the hand regions are closer to the camera rather than the face region (Fig. 3). Furthermore, the depth information is used to resolve complex background (i.e. neutralize complex background to increase the accuracy of skin segmentation for region of interest) completely.

Fig. 3. Solving overlapping problem between hand and face using depth map. (a) 2D image in which the face and the left hand are occluded. (b) 2D image with labeled hands and face without occlusion.

B. Tracking

In our system, a robust method for hand tracking is proposed using Mean-shift analysis in conjunction with depth map to retrieve the extracted features during occlusion. Mean-shift analysis uses the gradient of Bhattacharyya coefficient [18] as a similarity function to derive the candidate of the hand which is mostly similar to a given hand target model. This structure correctly extracts a set of hand postures to track the hand motion. The motivation behind mean shift analysis is to achieve accurate and robust hand tracking.

C. Feature Extraction

Selection of good features for the recognition of the hand gesture path plays a significant role in system performance. There are three basic features: location, orientation and velocity. A gesture path is spatiotemporal pattern that consists of hand centroid points (x_{hand}, y_{hand}). Thus, the coordinates in the Cartesian space can be extracted from gesture frames directly. 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

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

of feature vectors, which are projected and clustered in space dimension to obtain discrete code words which are employed as an input to HMMs. This is done using *k*-means clustering algorithm [19], [20], [21], [22], which classifies the gesture pattern into *K* clusters in the feature space. *K* depends on the numbers of segmented parts in gesture numbers from 0 to 9; however, each straight-line segment is classified into a single cluster. Fig. 4 shows the Cluster trajectories of gesture path '3' and '5', which are projected according to location, orientation and velocity features. It is noted that the cluster trajectories for gesture paths '3' and '5' nearly have the same cluster indices in some parts.

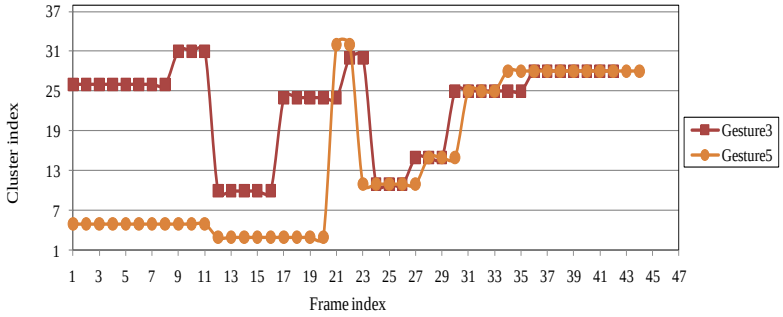

Fig. 4. Cluster trajectories for gesture path '3' and gesture path '5'.

D. Classification via HMMs

The main contribution of this paper is to propose a forward gesture spotting scheme which handles hand gesture spotting and recognition of numbers (0-9) simultaneously. This scheme uses a stochastic method for designing a non-gesture model by HMMs models with no training data. Furthermore, this scheme solves the issues of time delay between the spotting and the recognition task. The following sections describe how HMMs is used for hand gesture spotting and recognition. In addition, how to model gesture patterns discriminately and how to model non-gesture patterns effectively without training data for non-gesture patterns. HMMs are capable of modeling spatio-temporal time series of gestures effectively and can handle non-gesture patterns (garbage model or filler model). To proposed the non-gesture model provides a confidence measure based on the calculated likelihood of gesture models which is used as an adaptive threshold to find the start and the end points of key gestures which are embedded in the input video sequences. The performance of non-gesture model is improved using relative entropy measure to alleviate the problem of increasing number of states [23] (Fig. 5).

1. Gesture Model

For each reference gesture, each HMM state represents its local segmental part. However, the transition among states represents the sequential order structure in a gesture path. The number of HMMs states is an important parameter for each reference gesture. When the number of training data samples is insufficient, the use of excessive state numbers causes the over-fitting problem.

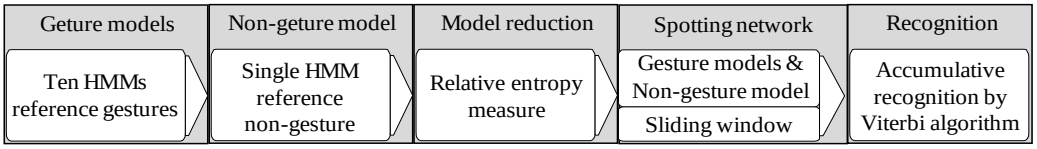

Fig. 5. Road map of gesture spotting and recognition using HMMs

In addition, the discrimination power of HMMs is decreased in case of using insufficient number of states, because more than one segmented part of graphical pattern is modeled on one state. Moreover, the number of states in our gesture spotting system is based on the complexity of each gesture number and is determined by mapping each straight-line segment into a single HMM state (Fig. 6). In practice, the Left-right Banded (LRB) model is considered for the following reasons. Since each state in Ergodic topology has many transitions than LR (Left-right) and LRB topologies, the structure data can be lost easily. On the other hand, LRB topology has no backward transition where the state index either increases or stays the same as time increases. In addition, LRB topology is more restricted than LR topology and simple for training data and is able to match the data with the

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

model. Therefore, Baum-Welch (BW) algorithm plays a significant role in our system, where it is used to do a full training for the initialized HMMs parameters $\lambda = (\pi, A, B)$. For more details, the reader can refer [12].

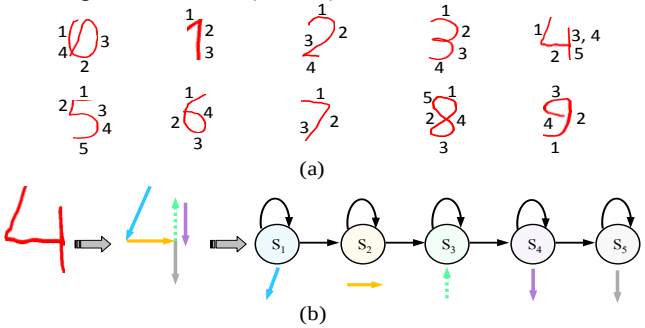

Fig. 6. The hand gesture paths and straight-line segmentation. (a) The gesture paths from hand motion trajectory for numbers (0-9) with its segmented parts. (b) The LRB topology with segmented line for a gesture path '4'.

2. Non-gesture Model

It is not easy to obtain the set of non-gesture patterns because there are infinite varieties of meaningless motion. So, all other patterns other than reference patterns are modeled by a single HMM called a non-gesture model (garbage model) [11], [24]. According to the property of HMM's internal segmentation, the self-transition for each state represents a line-segmented pattern of a gesture path and the outgoing transition from states lead to the rest of sequential segmented patterns in a gesture. Using this property, a model so-called Ergodic is created in which its states are copied from all gesture references in the system and then fully connect these states. The non-gesture model is constructed by copying the states of all gesture models in the system as follows (Fig. 7):

1) Copy all states from all gesture models, each with an output observation probability *bj* (*m*).
Then, re-estimate the probabilities with Gaussian distribution smoothing filter to make the states represent any pattern. After that, the floor smoothing is applied.

$$non-gesture(b_j(m)) = \frac{1}{\sqrt{2\pi}}.\exp\left(\frac{(b_j(m))^2}{2\sigma^2}\right) \quad (1)$$

2) The probabilities of self-transitions are copied as in the gesture models because each state represents a primitive unit (i.e. segmented pattern) of a gesture. The number of these units constitutes the target gesture.
3) The probabilities of all outgoing transitions are calculated as follow:

$$\hat{a}_{ij} = \frac{1-a_{ij}}{N-1} \quad for\ all\ \ j;\ \ i \neq j \quad (2)$$

where $b_j(m)$ is the emission probability of state j at observation value m, $\hat{a}_{ij}$ represents the transition probabilities of non-gesture model from state s_i to state s_j , a_{ij} is the transition probabilities of gesture models from state s_i to state s_j and N is the number of states in all gesture models. Also, the likelihood of the non-gesture model gives a confidence measure for the calculated likelihood by other gesture models because a confidence measure is based on the differential probability value. This value represents the difference between the observations probability of maximal gestures models and non-gesture model for an input pattern. Thereby, the confidence measure is used as an adaptive threshold for choosing the desired gesture model or gesture spotting.

3. Model Reduction

The number of states in the non-gesture model is equal to the sum of all states for all gesture models except the two dummy states (Start state: ST; End state: ET). This means the number of states for non-gesture model increases as the number of gesture model increases. Furthermore, an increase in the number of states does not affect the recognition rate, but dues to a waste of time and space. To alleviate this problem, relative entropy [23] is used to reduce the non-gesture model states because there are many states with similar probability distribution. The discrimination of input pattern is computationally expensive when the number of states for non-gesture model is

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

increased. The main advantage of using relative entropy is to reduce the number of states which constitutes the non-gesture model. Thus, the speed of computational process is increased as well as reducing the time and space.

4. Gesture Spotting Network

In continuous hand motion, key gestures appear intermittently with pre- and post-gestures (i.e. transition for connecting key gestures). To spot these key gestures, gesture spotting network is constructed as shown in Fig. 7.

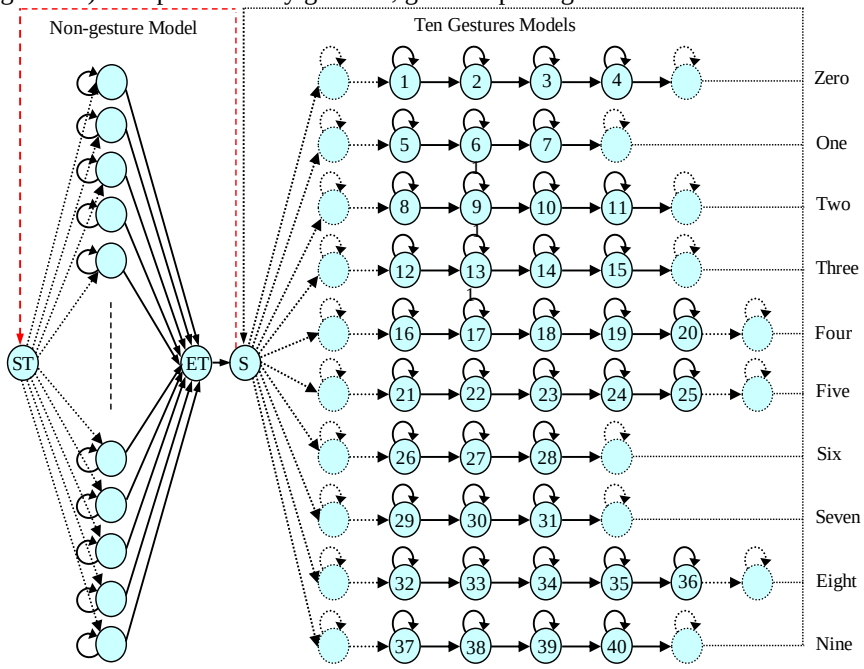

Fig. 7. The gesture spotting network which contains ten number gesture models from 0 to 9 and are designed by using LRB model with varying states from 3 to 5 and the Non-gesture model.

Moreover, the gesture spotting network can be easily expanded the vocabularies by adding a new key gesture HMM model and then rebuilding a non-gesture model. This network contains ten gesture models for numbers from 0 to 9. These ten models are designed using LRB model with number of states ranging from 3 to 5 based on its complexity. Additionally, it also contains non-gesture model after states reduction by relative entropy measure and the dummy start state S. The gesture spotting network finds the start and the end points of key gestures which are embedded in the input video stream and performs the segmentation and the recognition tasks simultaneously.

5. Spotting and Recognition

For forward spotting, a differential probability (*DP*) value is defined by the difference between the observation probability value of maximal gesture models and non-gesture model (Fig. 8). The maximal gesture is defined as a gesture having the largest value among all ten gestures P ($O|\lambda_g$) (g is the index of gesture models from 0 to 9). The transition from non-gesture to gesture occurs when the *DP* value changes from negative to positive (Eq. 3, where O is possibly as gesture g). Similarly, the transition from gesture to non-gesture occurs at the time when the *DP* value changes from positive to negative (Eq. 4, where O cannot be a gesture). Consequently, these observations are employed as a rule to detect the start and the end point of gestures. Here, the *DP* value represents an adaptive threshold which is used for selecting the desired gesture model or gesture spotting.

$$\exists g : P(O \mid \lambda g) > P(O|\lambda_{non\_gesture}) \quad (3)$$

$$\forall g : P(O \mid \lambda g) < P(O|\lambda_{non\_gesture}) \quad (4)$$

The proposed gesture spotting system contains two main modules: segmentation module (segmentation module is also called spotting module) and recognition module.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

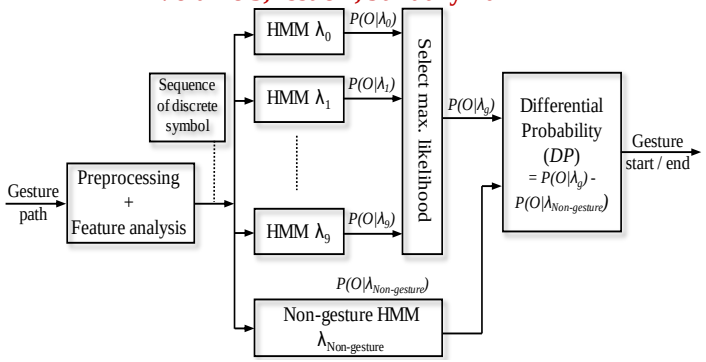

Fig. 8. Simplified structure showing the main modules for hand gesture spotting via HMMs, where the start and end points are based on differential probability value.

In gesture segmentation module, a sliding window technique is used. This technique calculates the observation probability of all gesture models and non-gesture model for observed segmented Parts to spot the start point by *DP* value. The sliding window (*Sw*) contains a number of sequential observations instead of a single observation (Fig.9). It is used to reduce the impact of observation changes for a short interval which are caused by incomplete feature extraction. The optimal value of sliding window is empirically determined with value 5 where the system is the best in term of results. The gesture recognition module is activated after detecting the start point from continuous image sequences. The main objective is to perform the recognition process accumulatively for the segmented parts until it receives the end signal of key gesture. Therefore, the type of observed gesture segmentation (arg max $P(O \mid \lambda_g)$ is decided at this point using Viterbi algorithm. Then, the processes of these modules are iterated until no more input stream of gesture images exist. Fig. 9 illustrates the work of sliding window and the recognition of observed sequences accumulatively. Assume that, the size of sliding window is S_w and the input observation sequence with length *T is* $O = \{o_1, o_2, ..., o_t, ..., o_T\}$. Firstly, the window size is initialized with an observation sequence $O_t = 0 = \{o_1, o_2, ..., o_{Sw}\}$. The *DP* value is equal to difference observation probability between the maximal gesture and the Non-gesture as follows:

$$DP(t) = \max_g P(O_t \mid \lambda_g) - P(O_t \mid \lambda_{Non\_gesture}) \quad (5)$$

When the value of *DP* (*t*) is negative, the start point in this case is not detected and therefore the sliding window is shifted on unit (i.e. $O_{t+1} = \{o_{t+1}, o_{t+2}, ..., o_{Sw+t}\}$). This process is repeated until *DP* value is positive. In the case of *DP* value is positive; assume that O'_1 represents the first partial key gesture segmented. Then, the observed key gesture segmented is represented by union of all possible partial gesture segments $O' = \{O'_1 \cup O'_2 \cup ...\}$. At each step, the gesture type of O' is determined. When the value of *DP* becomes negative again or there is no gesture image, the final gesture type *g* of observed gesture segment O' is determined by Viterbi algorithm. When there are more gesture images, the previous steps are repeated with re-initializing the sliding window at the next time *t*. Thus, the use of HMMs in conjunction with a relative entropy measure and sliding window scheme are capable of modeling spatio-temporal time series of gestures as well as handling non-gesture patterns and resolve the issues of time delay between gesture spotting and recognition.

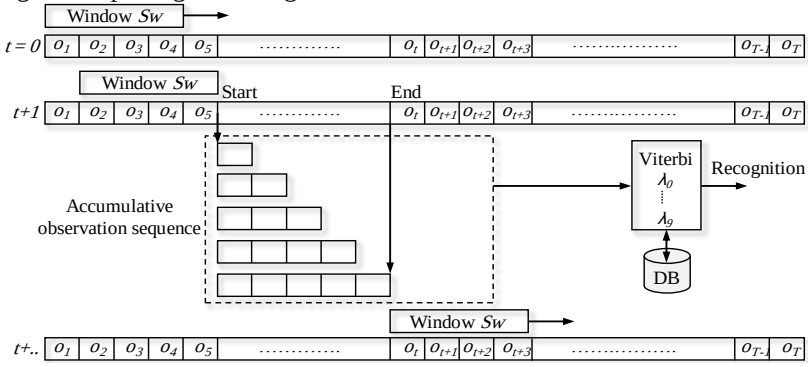

Fig. 9. Block diagram shows the work of sliding window. The Viterbi algorithm recognizes the segmented parts after detecting the start point

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

III. RESULTS AND DISCUSSION

The input images were captured by Bumblebee stereo camera system which has 6 mm focal length at 15FPS with 240 × 320 pixels image resolution, Matlab and C++ language implementation. Classification results are based on our database and it contains 600 video samples for isolated gestures which are captured from three persons on a set of numbers. Each number from 0 to 9 was based on 42 videos for training and 18 video samples for testing (In total, 420 video samples for training and 180 video samples for testing). Also, the database contains 280 video samples of continuous hand motion for testing. Each video sample either contains one or more meaningful gestures. The HMMs have been trained by BW algorithm. The inference (i.e. recognition) process uses forward score of each sample to select the model with the highest likelihood.

The gesture recognition module matches the tested gesture against database of reference gestures to classify which class it belongs to. The higher priority was computed by Viterbi algorithm to recognize the numbers frame by frame using LRB topology with different number of states ranging from 3 to 5 based on their complexity. Fig. 10 illustrates the result of isolated gesture '3' with high four priorities while the probability of Non-gesture model before and after state reduction is the same. Moreover, the number of states of Non-gesture model before state reduction is 40 and after reduction is 22 states. This in turn leads to several advantages such as saving time and space and most importantly, makes the system appropriate to real-time applications. From table 1, the recognition ratio of isolated gestures achieves best results with 97.78%. The recognition ratio is the number of correctly (true) recognized gestures to the number of tested gestures (Eq. 6).

$$\text{Recognition} = \frac{no.of\ recognized\ gestures}{no.of\ tested\ gestures} \times 100 \tag{6}$$

In automatic gesture spotting task, there are three types of errors called insertion (I), substitution (S) and deletion (D). The insertion error is occurred when the spotter detects a nonexistent gesture. It is because the emission probability of the current state for a given observation sequence is equal to zero. A substitution error occurs when the key gesture is classified falsely (i.e. classifies the gesture as another gesture). This error is usually happened when the extracted features are falsely quantized to other code words. The deletion error happens when the spotter fails to detect a key gesture. In order to calculate the recognition ratio (Eq. 6), insertion errors are totally not considered. However, insertion errors are probably caused due to substitution and deletion errors because they are often considered as strong decision in determining the end point of gestures to eliminate all or part of the meaningful gestures from observation. Deletion errors directly affect the recognition ratio whereas insertion errors do not. However, the insertion errors affect the gesture spotting ratio directly. To take into consideration the effect of insertion errors, another performance measure called reliability is proposed by the following equation:

$$\text{Reliability} = \frac{no.of\ recognized\ gestures}{no.of\ tested\ gestures + no.of\ inseration\ errors} \times 100 \tag{7}$$

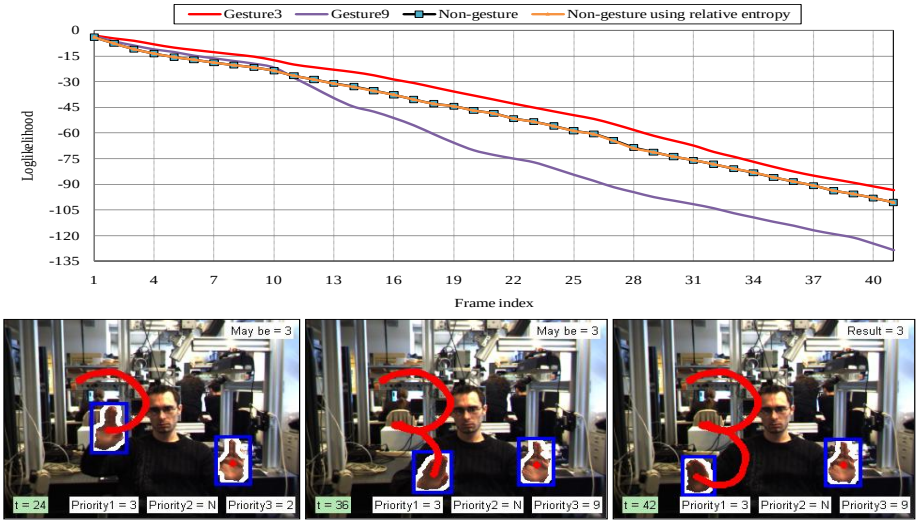

Fig. 10. Temporal evolution of four higher probabilities of the gestures '3', '9', 'Non-gesture' before and after state reduction. The probability of Non-gesture model before and after state reduction is the same. In the image sequences, the high priority is gesture '3' and the second priority refers to Non-gesture 'N' at t = 24. The final result is gesture number '3' at t = 42.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

TABLE 1: Isolated gesture recognition and key spotting gesture results for gesture numbers from '0' to '9' using HMMs at sliding window equal to 5.

Gesture path	Train data	Isolated gestures results			Key gestures spotting results					
		Test	Correct	Recognition (%)	Test	I	D	S	Correct	Reliability (%)
'0'	42	18	17	94.44	28	2	1	2	25	83.33
'1'	42	18	18	100.00	28	0	1	1	26	92.86
'2'	42	18	17	94.44	28	0	0	2	26	92.86
'3'	42	18	18	100.00	28	0	0	0	28	100.00
'4'	42	18	18	100.00	28	0	0	1	27	96.43
'5'	42	18	18	100.00	28	0	1	1	26	92.86
'6'	42	18	17	94.44	28	1	1	1	26	89.66
'7'	42	18	18	100.00	28	0	0	0	28	100.00
'8'	42	18	17	94.44	28	1	0	2	26	89.66
'9'	42	18	18	100.00	28	0	1	0	27	96.43
Total	420	180	176	97.78	280	4	5	10	265	93.31

The recognition ratio and the reliability are computed based on the number of spotting errors (Table2). The gesture spotting accuracy is measured according to different sliding window size ranging from 1 to 8. Furthermore, the gesture spotting accuracy is improved initially as the sliding window size increase, but degrades as sliding window size increase further. Therefore, the optimal size of sliding window is 5 empirically where the reliability of automatic gesture spotting system achieves 93.31%. From Table 2, the number of errors decreases sharply between *Sw* =1 and *Sw* = 4. However, deletion, insertion and substitute errors begin to increase after *Sw*=4. The increase in the size of *Sw* means that it contains some of observation features belong to gesture patterns and others belong to non-gesture patterns, and hence this leads to loss of starting and ending points of meaningful gestures. Furthermore, the yield of isolated training data is higher than isolated testing data. In addition, the overall recognition rate is the average of training and testing recognition rate and achieved 98.35% at the sliding window size equal to 5. Fig. 11 shows the results of continuous gesture path which contains one meaningful gestures '6' where the start point at frame index = 15 and the end point at frame index = 50. Moreover, the proposed system automatically recognizes isolated and key hand gestures with superior performance and low computational complexity when applied to several video samples containing confusing situations such as occlusion between hands and face. Experimental results of HMMs show that the proposed system automatically recognizes isolated gestures with 97.78% and key gestures with 93.31% reliability. It is noted that the proposed system achieved high recognition rate for isolated gestures and is due to a good election for the set of feature candidates to optimally discriminate among input patterns. Also, a careful experimental is based on selection of initialization parameters for the training process. In addition, HMMs have the ability to efficiently alleviate spatiotemporal variabilities. Thus, this system is capable for real-time applications and resolves the issues of time delay between spotting and recognition tasks.

TABLE 2: Results of isolated gestures recognition and key gestures spotting with different size of sliding window (*Sw*) ranging from 1 to 8 via HMMs.

Sw	Train data	Isolated gestures results				Spotting key gestures results					
		Test data	Recognition (%)			Test data	Error types			Spotting (%)	
			Train	Test	Overall		I	D	S	Rec.	Rel.
1	420	180	86.79	86.11	86.45	280	11	22	34	80.00	76.98
2	420	180	89.29	87.78	88.53	280	8	19	31	82.14	79.86
3	420	180	92.50	91.11	91.81	280	5	9	16	91.07	89.47
4	420	180	95.00	92.22	93.61	280	4	9	16	91.07	89.79
5	420	180	98.93	97.78	98.35	280	4	5	10	94.64	93.31

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

6	420	180	96.07	93.89	94.98	280	4	8	13	92.50	91.20
7	420	180	95.71	94.44	95.08	280	5	7	14	92.50	90.88
8	420	180	95.35	94.44	94.90	280	6	7	17	91.43	89.51

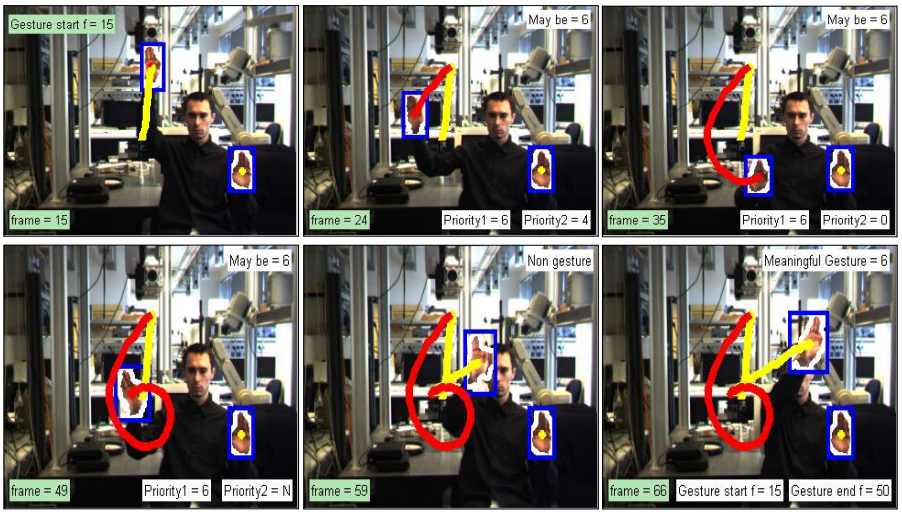

Fig. 11. Image sequences contain one meaningful gesture '6', where the start point at frame 15 and the end point at frame 50. 'N' refers to Non-gesture.

V. CONCLUSION

This paper proposes robust methods for hand gesture spotting and recognition using HMMs. Our methods perform the hand gesture spotting and recognition tasks simultaneously. Furthermore, they are suitable for real-time applications and solve the issues of time delay between the spotting and the recognition tasks. The results show that; the proposed methods can successfully spot and recognize meaningful gestures numbers (0-9) with 93.31%.. In addition, the model inference by HMMs is faster and the saving time is 66.42% using relative entropy. In future, we expect an ongoing increase of work on extending the proposed method with CRFs to Latent-Dynamic Conditional Random Fields. This builds on try-and-error to optimize LDCRFs parameters.

REFERENCES

[1] T. Starner, J. Weaver, and A. Pentland, Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video, IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. 20, No. 12, pp. 1371-1375, 1998.

[2] H. Yang, A. Park, and S. Lee, Gesture Spotting and Recognition for Human- Robot Interaction, IEEE Transaction on Robotics, Vol. 23, No. 2, pp. 256-270, 2007.

[3] W. Freeman and M. Roth, Orientation Histograms for Hand Gesture Recognition, International Workshop on Automatic Face and Gesture Recognition, pp. 296-301, 1995.

[4] S. Ju, M. Black, S. Minneman, and D. Kimber, Analysis of Gesture and Action in Technical Talks for Video Indexing, IEEE Conference on Computer Vision and Pattern Recognition, pp. 595-601, 1997.

[5] V. Nair and J. J. Clark, Automated Visual Surveillance Using Hidden Markov Models, Conference of Vision Interface, pp.88-92, 2002.

[6] R. C. Rose, Discriminant Word spotting Techniques for Rejection Non-vocabulary utterances in Unconstrained Speech, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 2, pp. 105-108, 1992.

[7] F. R. Chen, L. D. Wilcox, and D. S. Bloomberg, Word Spotting in Scanned Images using Hidden Markov Models, IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, pp. 1-4, 1993.

ISSN: 2319-5967
ISO 9001:2008 Certified
International Journal of Engineering Science and Innovative Technology (IJESIT)
Volume 3, Issue 1, January 2014

[8] K. Takahashi, S. Seki, , and R. Oka, Spotting Recognition of Human Gestures from Motion Images, Technical Report IE92-134, pp. 9-16, 1992.

[9] T. Baudel and M. Beaudouin, CHARADE: Remote Control of Objects using Free-Hand Gestures, Communications of ACM, Vol. 36, No. 7, pp. 28-35, 1993.

[10] A. Wexelblat, Natural Gesture in Virtual Environments, Proceedings of Virtual Reality Software and Technology Conference, pp. 5-16, 1994.

[11] H. Lee and J. Kim, An HMM-Based Threshold Model Approach for Gesture Recognition, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol. 21, No. 10, pp. 961-973, 1999.

[12] 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, Journal of WSCG, Vol.16, No. 1, ISSN: 1213-6972, pp. 65-72, 2008.

[13] S. L. Phung, A. Bouzerdoum, and D. Chai, A Novel Skin Color Model in YCbCr Color Space and its Application to Human Face Detection, IEEE International Conference on Image Processing, pp. 289-292, 2002.

[14] M. Elmezain, A. Al-Hamadi, J. Appenrodt, and B. Michaelis, A Hidden Markov Model-Based Isolated and Meaningful Hand Gesture Recognition, International Journal of Electrical, Computer, and Systems Engineering, Vol. 3, No. 3, ISSN:2070-3813, pp. 156- 163, 2009.

[15] S. Askar, Y. Kondratyuk, K. Elazouzi, P. Kauff, and O. Schreer, Vision-Based Skin-Color Segmentation of Moving Hands for Real- Time Application, 1st European on CVMP, pp. 79-85, 2004.

[16] M. Elmezain, A. Al-Hamadi, O. Rashid, and B. Michaelis, Posture and Gesture Recognition for Human-Computer Interaction, In-Tech Olajnica 19/2, Croatia, "Advanced Technologies", ISBN: 978-953-307-009-4, pp. 415-440, 2009.

[17] R. Niese, A. Al-Hamadi, and B. Michaelis, A Novel Method for 3D Face Detection and Normalization, Journal of Multimedia, ISSN: 1796-2048, Vol. 2, No. 5, pp. 1-12, 2007.

[18] D. Comaniciu, V. Ramesh, and P. Meer, Kernel-Based Object Tracking, IEEE Trans. on TPAM, pp. 564-577, 2003.

[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 TPAMI, Vol. 24, pp. 881-892, 2002.

[20] M. Elmezain, A. Al-Hamadi, S. Pathan, and B. Michaelis, Spatio-Temporal Feature Extraction-Based Hand Gesture Recognition for Isolated American Sign Language and Arabic Numbers, IEEE International Symposium on Image and Signal Processing and Analysis, pp. 254-259, 2009.

[21] M. Elmezain, A. Al-Hamadi, and B. Michaelis, Hand Gesture Recognition Based on Combined Features Extraction, International Conference on Machine Vision, Image Processing, and Pattern Analysis, PWASET, Vol. 60, pp. 459-464, 2009.

[22] M. Elmezain, A. Al-Hamadi, and B. Michaelis, Improving Hand Gesture Recognition using 3D Combined features, International Conference on Machine Vision, pp. 128-132, 2009.

[23] T. M. Cover and J. A. Thomas, Entropy, Relative Entropy and Mutual Information, Elements of Information Theory, pp. 12-49, 1991.

[24] M. Elmezain, A. Al-Hamadi, and B. Michaelis, A Novel System for Automatic Hand Gesture Spotting and Recognition in Stereo Color Image Sequences, Journal of WSCG, Vol.17, No. 1, ISSN:1213-6972, pp. 89-96, 2009.