Journal of Software Engineering and Applications, 2015, 8, 313-323
Published Online July 2015 in SciRes. http://www.scirp.org/journal/jsea
http://dx.doi.org/10.4236/jsea.2015.87032

Vision-Based Hand Gesture Spotting and Recognition Using CRF and SVM

Fayed F. M. Ghaleb[1], Ebrahim A. Youness[2], Mahmoud Elmezain[2], Fatma Sh. Dewdar[2]

[1]Faculty of Science, Mathematices Department, Ain Shams University, Cairo, Egypt
[2]Faculty of Science, Computer Science Division, Tanta University, Tanta, Egypt
Email: Mahmoud.Elmezain@tuscs.com

Received 27 May 2015; accepted 21 July 2015; published 24 July 2015

Abstract

In this paper, a novel gesture spotting and recognition technique is proposed to handle hand gesture from continuous hand motion based on Conditional Random Fields in conjunction with Support Vector Machine. Firstly, YC_bC_r color space and 3D depth map are used to detect and segment the hand. The depth map is to neutralize complex background sense. Secondly, 3D spatio-temporal features for hand volume of dynamic affine-invariants like elliptic Fourier and Zernike moments are extracted, in addition to three orientations motion features. Finally, the hand gesture is spotted and recognized by using the discriminative Conditional Random Fields Model. Accordingly, a Support Vector Machine verifies the hand shape at the start and the end point of meaningful gesture, which enforces vigorous view invariant task. Experiments demonstrate that the proposed method can successfully spot and recognize hand gesture from continuous hand motion data with 92.50% recognition rate.

Keywords

Human Computer Interaction, Conditional Random Fields, Support Vector Machine, Elliptic Fourier, Zernike Moments

1. Introduction

The task of locating the start and the end points that correspond to a gesture of interest is a challenging task in Human Computer Interaction. We define a gesture as the motion of the hand to communicate with a computer. The task of locating meaningful patterns from input signals is called pattern spotting [1] [2]. In gesture spotting, an instance of pattern spotting, it is required to locate the start point and the end point of a gesture. The gesture spotting has two major difficulties: segmentations [3] [4] and spatio-temporal variabilities [5] [6]. The segmen-

How to cite this paper: Ghaleb, F.F.M., Youness, E.A., Elmezain, M. and Dewdar, F.S. (2015) Vision-Based Hand Gesture Spotting and Recognition Using CRF and SVM. *Journal of Software Engineering and Applications*, **8**, 313-323.
http://dx.doi.org/10.4236/jsea.2015.87032

tation problem is how to determine when a gesture starts and when it ends in a continuous hand trajectory. As the gesturer switches from one gesture to another, his hand makes an intermediate move linking the two gestures. A gesture recognizer may attempt to recognize this inevitable intermediate motion as a meaningful one. Without segmentation, the recognizer should try to match reference patterns with all possible segments of input signals. The other difficulties of gesture spotting are that the same gesture varies dynamically in shape and duration; even of the same gesturer. Therefore, the recognizer should consider both the spatial and the temporal variabilities simultaneously. An ideal recognizer will extract gesture segments from the input signal, and match them with reference patterns regardless of the spatio-temporal variabilities. Tracking methods that depend entirely on an image-plane representation of the hand have been worked on extensively. Typically such systems are computationally less expensive than those methods that use a 3D model. Moment invariants, as discriminative feature descriptors, have been used for shape representation for many years. The shape-based image invariants can be divided into two different categories: boundary based image invariants such as Fourier descriptors; region-based image invariants included various moment-based invariants such as Zernike moments. There are two types of shape-based image invariants: boundary-based and region-based. The boundary based image invariants focus on the properties contained in the image's contour while the region-based image invariants take the whole image area as the research object.

Boundary-based invariants such as Fourier descriptors explore only the contour information; they cannot capture the interior content of the shape. On the other hand, these methods cannot deal with disjoint shapes where single closed boundary may not be available; therefore, they have limited applications. For region-based invariants, all of the pixels of the image are taken into account to represent the shape. Because region-based invariants combine information of an entire image region rather than exploiting information just along the boundary pixels, they can capture more information from the image. The region-based invariants can also be employed to describe disjoint shapes.

Lee, H.-K. and Kim, J.H. [7] develope a new method by using the Hidden Markov Model based technique. To handle non-gesture patterns, they introduce the concept of a threshold model that calculates the likelihood threshold of an input pattern and provides a confirmation mechanism for the provisionally matched gesture patterns. *Yang, H.-D., Sclaroff, S. and Lee, S.-W.* [8] propose a novel method for designing threshold models in a conditional random field (CRF) model which performs an adaptive threshold for distinguishing between signs in a vocabulary and non-sign patterns. These methods have the following consequent drawback to detect the reliable end point of a gesture and find the start point by back-tracking. Therefore, the delayed response may cause one to wonder whether one's gesture has been recognized correctly or not. For that reason, the systems in [7] [8] are not capable for real-time applications.

To face the mentioned challenges, CRF forward gesture spotting by using Circular Buffer method is proposed, which simultaneously handles the hand gesture spotting and recognition in stereo color image sequences without time delay. A hand appearance-based sign verification method using SVM is considered to further improve sign language spotting accuracy. Additionally, a depth image sequence is exploited 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. Our experiments on own dataset, showed that the proposed approach is more robust and yields promising results when comparing favorably with those previously reported throughout the literature.

2. Hand Gesture Spotting and Recognition Approach

An application of gesture-based interaction with Arabic numbers (0 - 9) is implemented to demonstrate the coaction of suggested components and the effectiveness of gesture spotting & recognition approach (**Figure 1**).

2.1. Preprocessing

Automatic segmentation and preprocessing is an important stage in our approach. The segmentation of the hand takes place using color information and 3D depth map. Firstly, the hand is segmented (*i.e.* Area of interest (AOI)) using Gaussian Mixture Model (GMM) over YC_bC_r color space, where Y channel represents brightness and (C_b, C_r) channels refer to chrominance [9]. Channel Y is eliminated to reduce the effect of brightness variation and only the chrominance channels are employed that fully represent the color information. Accordingly, Image acquisition includes 2D image sequences and depth image sequences, which devised by an algorithm of passive stereo measuring based on mean absolute difference and the known calibration data of the cameras. The

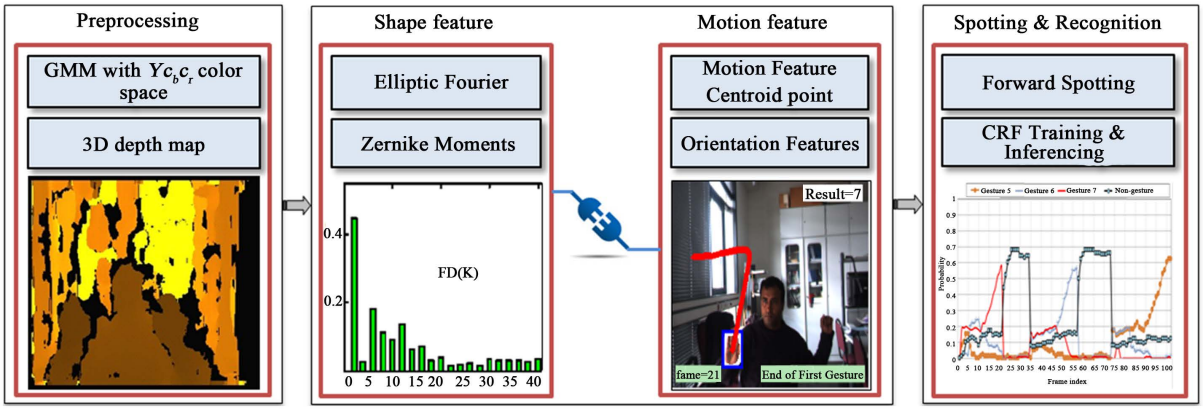


Figure 1. Hand gesture spotting and recognition concept.

depth map is to neutralize complex senses background to increase the robustness of skin segmentation for AOI completely (**Figure 2**).

2.2. Tracking and Feature Extraction

To retrieve the extracted features during occlusion, a robust method for hand tracking is considered using Mean-shift analysis in conjunction with depth map. The motivation behind mean shift analysis is to achieve accurate and robust hand tracking. Mean-shift analysis uses the gradient of Bhattacharyya coefficient [10] 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. After that, two types of feature are employed to correctly spot and recognize hand gesture. The orientation features of hand image sequences are extracted by the trajectory of the hand motion centroid. Additionally, the shape features are considered with a variety of invariant descriptors such as elliptic Fourier's descriptors and invariant Zernike moments.

2.2.1. Orientation Features

The orientation gives the direction of the hand when traverses in space during the gesture making process. A gesture path is spatio-temporal pattern that consists of centroid points (xhand, yhand). Therefore, orientation feature is based on; the calculation of the hand displacement vector at every point which 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}).

$$\theta_{1t} = \tan^{-1}\left(\frac{y_{t+1}-c_y}{x_{t+1}-c_x}\right),\ \theta_{2t} = \tan^{-1}\left(\frac{y_{t+1}-y_t}{x_{t+1}-x_t}\right),\ \theta_{3t} = \tan^{-1}\left(\frac{y_{t+1}-y_1}{x_{t+1}-x_1}\right) \tag{2.1}$$

$$\left(c_x, c_y\right) = \frac{1}{n}\left(\sum_{t=1}^{n} x_t, \sum_{t=1}^{n} y_t\right) \tag{2.2}$$

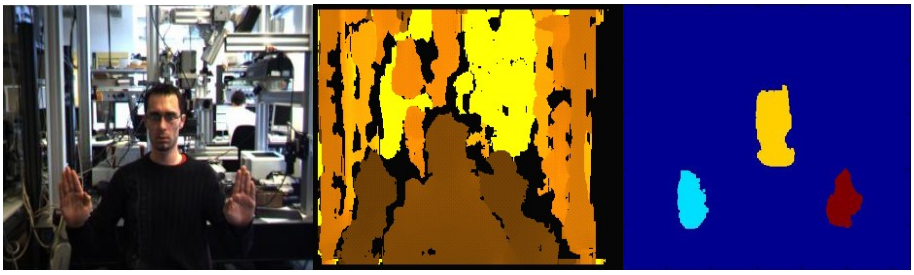
where (c_x, c_y) refers to the centroid of gravity at n points, and T represents the length of hand gesture path such that $t = 1, 2, \cdots, T-1$.

In this manner, gesture is represented as an ordered sequence of feature vectors, which are projected and clustered in space dimension to obtain discrete code words. This is done using k-means clustering algorithm [11], which classifies the gesture pattern into K clusters in the feature space.

2.2.2. SVM-Based Dynamic Affine-Invariants Features

The shape flow as the global flow of hand is characterized and stated by the elliptic Fourier descriptors and Zernike moments $G_i = \left[C_{xk}, C_{yk}, z_{00}, z_{11}, z_{22}\right]^{\mathrm{T}}$ that described as follows:

The elliptic Fourier descriptors for action silhouettes are obtained using a trigonometric form of the shape curve C_k.



(a) (b) (c)

Figure 2. (a) Source image frame; (b) Depth value from the Bumblebee stereo camera system; (c) Skin color segmentation of hands.

$$C_k = C_{xk} + jC_{yk} \tag{2.3}$$

where

$$C_{xk} = \frac{1}{T}\int_0^T x(t)\mathrm{e}^{-jk\omega t},\ C_{yk} = \frac{1}{T}\int_0^T y(t)\mathrm{e}^{-jk\omega t} \tag{2.4}$$

ω defines the fundament frequency and is equal to $T/2\pi$. T referents the function period and k refers to a harmonic number. It can be verified that this choice of coefficients guarantees that the resulting curve descriptors are invariant to shape translation, rotation and scaling, and they are independent from the choice of starting point on a contour [12].

Invariance of hand image can be achieved by using Zernike moments, which give an orthogonal set of rotation-invariant moments. Additionally, scale and translation invariance can be implemented using moment normalization [13]. More simply, the complex Zernike moment (Z_{pq}) with an order p and repetition q of image intensity function $f(\rho,\theta)$ is:

$$Z_{pq} = \frac{p+1}{\delta_N}\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f\left(\frac{x}{a}+\bar{x},\frac{y}{a}+\bar{y},t\right)R_{pq}(\rho)\mathrm{e}^{-jq\theta} \tag{2.5}$$

Here, δ_N is a normalization factor, p is a positive integer while q either negative or positive integer subject to the constraints $p - |q|$ = even and $|q| \le p$. The function f is normalized with respect to scale and translation by using the center of silhouette image $(\bar{x},\bar{y})$ and the scale factored a. $R_{pq}(\rho)$ represents are dial polynomial [12]. Thus, the Zernike moment features invariant along with geometric features to shape translation, rotation and scaling with remarkable similarity to the Hu invariant moments is assigned by $G_z = [z_{00}, z_{11}, z_{22}]$.

2.3. Spotting and Recognition

To spot meaningful gestures of numbers (0 - 9), which are embedded in the input video stream accurately, a two-layer CRF architecture is applied (**Figure 3**). In the first layer a stochastic method for designing a non-gesture model with CRF is proposed without training data. CRF is capable of modeling spatio-temporal time series of gestures effectively and can handle non-gesture patterns. 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. As, a forward spotting technique in conjunction with a Differential Probability (DP) value and circular buffer is employed to discriminate between meaningful gestures and non-gesture patterns. In the second layer, CRF are used to find the maximal gesture, which having the largest value among all ten labels gestures. Finally, SVM decide the final decision based on the building dataset of hand shape with elliptic Fourier and Zernike moments.

2.3.1. Spotting with CRFs

CRFs are a framework based on conditional probability approaches for segmenting and labeling sequential data. CRFs use a single exponential distribution to model all labels of given observations. Therefore, there is a trade-off in the weights of each feature function, for each state. In our application, each state corresponds to segments of the number. In addition, each label in CRFs is employed as exponential model to conditional probabilities of

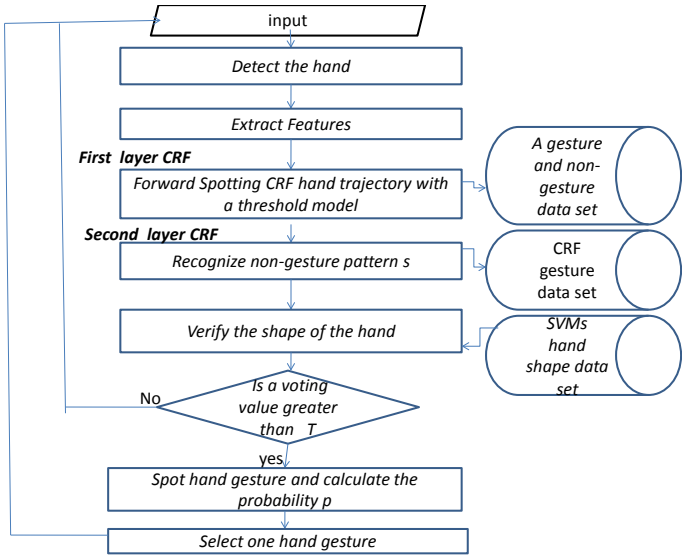


Figure 3. A two layer flow chart for spotting meaningful gestures based on CRFs and SVMs.

the next label for a given current label.

(1) Spotting with CRFs

Conditional Random Fields were developed for labeling sequential data (*i.e.* determining the probability of a given label sequence for a given input sequence) and are undirected graphical models (*i.e.* discriminative models). The structure of current label is design to form the chain with an edge between itself and previous label. Moreover, each label corresponds to a gesture number. The probability of label sequence y for a given observation sequence *O* is calculated as:

$$P\left(y|O,\theta\right)=\frac{1}{Z\left(O,\theta\right)}\cdot\exp\left(\sum_{i=1}^{n}F_{\theta}\left(y_{i-1},y_{i},O,i\right)\right) \tag{2.6}$$

where parameter $\theta=\left(\lambda_1,\lambda_2,\cdots,\lambda_{N_f};\mu_1,\mu_2,\cdots,\mu_{N_g}\right)$, the number of transition feature function represents by N_f, the number of state feature function represents by N_g and the length of observation sequence *O* is *n*.

F_θ is defined as:

$$F_{\theta}=\sum_{f}\lambda_{f}t_{f}\left(y_{i-1},y_{i},O,i\right)+\sum_{g}\mu_{g}s_{g}\left(y_{i},O,i\right) \tag{2.7}$$

where a transition feature function at position *i* and *i* − 1 is $t_f\left(y_{i-1},y_i,O,i\right)$ (*i.e.* represents the weight on the transition from label *i* to label *i* − 1 when the current observation is *O*). State feature function at position *i* is sg (y_i, *O*, *i*) (i.e. represents the weight on the label i when the current observation is *O*). λ_f and μ_g represent the weights of the transition and state feature functions, respectively. *Z*(*O*, θ) is the normalized factor and is calculated as follows:

$$Z\left(O,\theta\right)=\sum_{y}\exp\left(\sum_{i=1}^{n}F_{\theta}\left(y_{i-1},y_{i},O,i\right)\right) \tag{2.8}$$

The CRFs are initially built without label for non-gesture pattern; Because CRFs use a single model for the joint probability of the sequences $p(y|O, \theta)$.

Using gradient ascent with the BFGS optimization technique with 300 iterations is used for trained CRFs to achieve optimal convergence. Therefore, the labels of CRFs are $y=\left\{Y_0,Y_1,\cdots,Y_9\right\}$. To create the Non-gesture model (*N-CRFs*) using the weights of transition and state features function of initial CRFs all other patterns than gesture patterns are modeled by adding a label (*N*) for non-gesture patterns. Moreover, $y_N=\left\{Y_0,Y_1,\cdots,Y_9,Y_N\right\}$ are the labels of *N-CRFs*. The proposed *N-CRFs* model does not need non-gesture patterns for training and also can better spot gestures and non-gesture patterns.

(2) N-CRFs Model Parameters

The label of non-gesture pattern is created, by using the weight of state and transition feature function of the initialized CRFs model. There are two main parameters of CRFs named state feature function and transition feature function as in Equation (2.7). From the idea of Dugad *et al*. [14] who propose an adaptive threshold model based on the mean and the variance of sample, the weight of state feature function is computed as:

$$\mu_g(N) = \overline{\mu_g} + T_N\sqrt{\sigma_g} \tag{2.9}$$

where $\overline{\mu_g}$ is the mean of state feature functions of the labels of initial CRFs from Y_0 to Y_9 and σ_g represent the variance of the gth state feature functions T_N reflects the width of state features function in some way. The optimal value of T_N is 0.7 and is determined by multiple experiments which have been conducted with a range of values on a training data set.

It is difficult to spot and recognize short gestures because short gestures have fewer samples than long gestures. A challenging problem is caused by the fact that there is a quite bit of variability in the same gesture even for the same person. A short gesture detector is added to avoid this problem, where the weights of self-transition feature functions are increased as follows:

$$\lambda_f(Y_l, Y_l) = \begin{cases} \lambda_f(Y_l, Y_l) + \Psi_f(y_l), & \text{if Cond 1} \\ \lambda_f(Y_l, Y_l), & \text{otherwise} \end{cases} \tag{2.10}$$

Such that:

$$\Psi_f(y_l) = \frac{\left(\overline{N}_{frame} - \sigma_{N_{frame}}\right) - N_{frame}(Y_l)}{\max_l N_{frame}(Y_l)} \tag{2.11}$$

and

$$\text{Cond 1}: N_{frame}(Y_l) < \left(\overline{N}_{frame} - \sigma_{N_{frame}}\right) \tag{2.12}$$

where $N_{\text{frame}}(Y_l)$ is the average frame number of a gesture Y_l, $\sigma_{N_{frame}}$ represents the average frame number of all gestures from Y_0 *to* Y_9 and $\sigma_{N_{frame}}$ is the variance of them. $\psi_f(y_l)$ is additional weight of the gesture Y_l notable in case of a short length gesture.

The weight of the self-transition feature function of the label of non-gesture patterns is approximately assigned with the maximum weight of transition feature functions to initialize CRFs as follows:

$$\lambda_f(Y_N, Y_N) = \max_l \lambda_f(Y_l, Y_l) + \frac{\sum_{i=1}^{l}\sum_{g=1}^{N_g}\mu_g(Y_l)}{\overline{N}_{state\_feature}} \tag{2.13}$$

where $\overline{N}_{state\_feature}$ is the average number of transition feature functions in which the weight is greater than zero. As described above about the transition parameters of non-gesture model, a method is employed to compute the weights of transition feature functions between the labels of gesture models and the label of non-gesture patterns. Therefore, the weights of transition feature functions from the non-gesture label to other labels are computed by the following equation:

$$\lambda_f(Y_N, Y_i) = \frac{\lambda_f(Y_N, Y_N)}{l}, \quad \forall i \in \{1, 2, \cdots, l\} \tag{2.14}$$

Additionally, the weights of transition feature functions from the gesture labels to non-gesture label occurs by the given equation below:

$$\lambda_f(Y_i, Y_N) = \frac{\lambda_f(Y_i, Y_i)}{l}, \quad \forall i \in \{1, 2, \cdots, l\} \tag{2.15}$$

As a result, the *N-CRFs* model can better spot gestures and non-gesture patterns.

2.3.2. CRF Forward Spotting via Circular Buffer

Several samples of each gesture in the gesture vocabulary are stored in the database to be used in training the classifier. A gesture is stored as a sequence of (*x*, *y*, *t*) joint coordinates. In recognition mode, a circular buffer is used to temporarily store the real-time information. The circular buffer contains a number of sequential observations instead of a single observation. It is used to reduce the impact of observation changes for a short interval which are caused by incomplete feature extraction. The circular buffer is set to store 13 frames. The size of the circular buffer is chosen to be equal to the shortest gesture (*i.e.* gesture "1" represents a shortest gesture in our system). This ensures that at some stage during a continuous motion, the buffer will be completely filled with gesture data, and will not contain any transitional motion data. In addition, a maximum of one complete gesture can be stored in the buffer at one time.

Figure 4 illustrates the operation of the circular buffer in gesture segmentation. At $t = t_1$ the buffer contains transitional data and gesture data. At $t = t_2$, the buffer is filled entirely with data from Gesture 2, and contains no transitional motion data. 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.

Assume that, the size of circular buffer is initialized with the input observation sequence with length $T = 13$ as in our system, $O = \{o_1, o_2, \cdots, o_T\}$. The *DP* value is equal to difference observation probability between the maximal gesture labels and the Non-gesture label.

When the value of $DP(t)$ at time t is negative, the start point in this case is not detected and therefore the circular buffer is shifted on unit (*i.e.* $O_{t+1} = o_{t+1}, o_{t+2}, \cdots, o_{T+1}$). This process is repeated until *DP* value is positive. In the case of DP value is positive; assume that A1 represents the first partial key gesture segmented. Then, the observed key gesture segmented is represented by union of all possible partial gesture segments $A = \{A_1 \cup A_2 \cup \cdots\}$. At each step, the gesture type of *A* is determined. When the value of *DP* becomes negative again or there is no gesture image, the final gesture label type of observed gesture segment *A* is determined. When there is more gesture images, the previous steps are repeated with re-initializing the circular buffer at the next time *t*. Therefore, a forward scheme has the ability to resolve the issues of time delay between gesture spotting and recognition.

2.3.3. SVMs-Based Gesture Verification

The main motivation behind using SVMs-based gesture verification to decide whether or not to accept a gesture spotted. This helps to discriminate gestures which may have similar hand motions but different hand shapes. SVMs are trained with a set of images of a gesture hand shape according to the extracted features via elliptic Fourier and Zernike moments. Therefore, the histograms of gradient features are extracted for training SVM from training samples [15].

In our work, the hand shape at the start of the gesture is verified. Then the hand appearance is verified over a period of several frames. As a result, voting value over frames is used to consider whether to accept or reject the gesture. If voting value is greater than a specific threshold, which experimentally determined, then the candidate gesture is chosen be a meaningful gesture.

3. Experimental Results

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 implementation. Classification results are based on our database that contains 600 video samples for isolated gestures (*i.e.* 60 video samples for each gesture from 0 to 9) which are captured from three persons. Each isolated number from 0 to 9 was based on 42 videos for training CRF and SVMs. Additionally, the database contains 280 video samples of continuous hand motion for testing. Each video sample either contains one or more meaningful gestures.

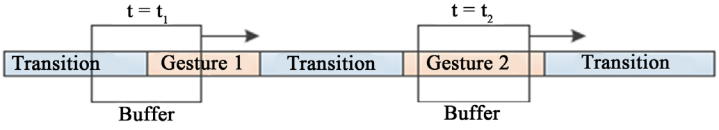


Figure 4. Gesture segmentation by circular buffer.

On a standard desktop PC, training process is more expensive for CRFs since the time which the model needs ranges from 20 minutes to several hours based on observation window. On the contrary, the inference (*i.e.* recognition) process is less costly and very fast for all models with sequences of several frames.

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 (Rec.) (Equation (3.2)), 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 (Rel.) is proposed by the following equation:

$$\text{Rel.} = \frac{\#\text{correctly recognized gestures}}{\#\text{test gestures} + \#\text{Inseration errors}} \times 100 \tag{3.1}$$

The recognition ratio and the reliability are computed based on the number of spotting errors (**Table 1**).

$$\text{Rec.} = \frac{\#\text{recognized gestures}}{\#\text{test gestures}} \times 100 \tag{3.2}$$

Experimental results of CRF show that the proposed method automatically recognizes meaningful gestures with 92.50% recognition (**Table 1**). It is noted that the proposed method achieved good recognition r ate due to a good election for the set of feature candidates to optimally discriminate among input patterns. In addition, A short gesture detector has the ability to efficiently alleviate spatio-temporal variability. Thus, this system is capable for real-time applications and resolves the issues of time delay between spotting and recognition tasks.

Lee, H.-K. and Kim, J.H. [7] developed a new method using the Hidden Markov Model based technique. To handle non-gesture patterns, they introduce the concept of a threshold model that calculates the likelihood threshold of an input pattern and provides a confirmation mechanism for the provisionally matched gesture patterns. For gesture segmentation, it detects the reliable end point of a gesture and finds the start point by back-tracking the Viterbi path from the end point (*i.e.* backward technique). The model performs gesture spotting with recognition

Table 1. Meaningful gesture spotting results for gesture numbers from “0” to “9”.

Gesture path	Train data	Meaningful gestures spotting results						
		Test	I	D	S	Correct	Rec. (%)	Rel. (%)
“0”	42	28	1	0	1	27	96.43	93.10
“1”	42	28	1	0	0	27	96.43	93.10
“2”	42	28	1	1	1	26	92.86	89.65
“3”	42	28	1	1	1	26	92.86	89.65
“4”	42	28	1	2	1	25	89.93	86.21
“5”	42	28	1	1	1	26	92.86	89.65
“6”	42	28	1	1	1	26	92.86	89.65
“7”	42	28	1	2	1	25	89.93	86.21
“8”	42	28	1	2	1	25	89.93	86.21
“9”	42	28	1	1	1	26	92.86	89.65
Total	420	280	10	11	9	259	92.50	89.31

rate 93.14%. However, the proposed method has a problem in that the system cannot report the detection of a gesture immediately after the system reaches its end point. It is because the endpoint detection process postpones the decision until the detection of the next gesture in order to avoid premature decision. The delayed response may cause one to wonder whether one's gesture has been recognized correctly or not. For that reason, their system is not capable for real-time applications.

Yang, H.-D., Sclaroff, S. and Lee, S.-W. [8] proposed a novel method for designing threshold models in a conditional random field (CRF) model which performs an adaptive threshold for distinguishing between signs in a vocabulary and non-sign patterns. A short-sign detector, a hand appearance-based sign verification method, and a sub-sign reasoning method are included to further improve sign language spotting accuracy. Their experiments demonstrate that their system can spot signs from continuous data with an 87.00% spotting rate. In order to spot signs from continuous data, a two-layer CRF architecture is applied with a sub-sign reasoning method, a short-sign detector, and an appearance-based shape verification method. In the first layer, in-vocabulary signs and non-sign patterns are discriminated by a threshold model with CRF. Spotted signs in the first layer are temporarily saved. Subsequent to detecting a sign sequence, the second layer CRF is applied to find sub-sign patterns. Sub-sign patterns within signs are modeled with a CRF. The input sequence of the second layer CRF includes more than one sign. The results of the second layer CRF are used to find the candidate label with the maximum probability. Finally, the appearance-based sign verification method is performed for the selected candidate sign. These steps are employed with end point detection technique that is not suitable for real-time applications.

Elmezain, M., Al-Hamadi, A. and Michaelis, B. [16] proposed a stochastic method for designing a non-gesture model with Hidden Markov Models (HMMs) versus Conditional Random Fields (CRFs). He used as an adaptive threshold to find the start and the end point of meaningful gestures, which are embedded in the input video stream. Also, he employed the forward spotting technique with sliding window with size ranging from 1 to 7 to empirically decide the optimal value. Their system was enhanced using relative entropy measure and increasing self-transition weight for HMMs and CRFs short gesture, respectively. The experiments of them could can successfully spot and recognize meaningful gestures with 93.31% and 90.49% reliability for HMMs and CRFs respectively.

In the light of this comparison (**Table 2**), it is being notice that the proposed approach performs competitively with another state-of-the-art method [7] [8] [16] in addition to carry out without sacrificing real-time performance. To assess the efficiency of the proposed method, the obtained results have been compared with those of other previously published studies in the literature, as shown in **Table 2**. From this comparison, it turns out that our approach performs competitively with other state-of-the-art approaches, and its results compared favorably with previously published results. Notably, all the methods that we compared our method with have used nearly similar experimental setups. Thus, the comparison is meaningful.

The image sequences depicted in **Figure 5** contain three key gestures "7", "6" and "5". The above graph of this figure considers only the temporal evolution of the probabilities of gestures "7", "6", "5" and non-gesture (for simplicity, the other curves are eliminated because their probabilities are low). The gesture "7" ends at frame index 21. Between frame index 22 and frame index 34, the highest priority is assigned to non-gesture label which means that the start point of second key gesture is not detected. At frame index 35, a new key gesture is started where the probability value of non-gesture label is not the highest value as compared to the other gesture labels. The gesture "6" ends at frame index 57. Between frame index 58 and frame index 73, the highest priority is assigned to non-gesture label. The gesture "5" starts at frame index 74 and ends at frame index 102.

Table 2. Comparison with the state-of-the-art.

Method	Recognition rate
Our method	92.50%
Lee, H.-K. and Kim, J. H. [7]	93.14%
Yang, H.-D., Sclaroff, S. and Lee, S.-W. [8]	87.00%
Elmezain, M. Al-Hamadi, A. and Michaelis, B. [16]	90.49%

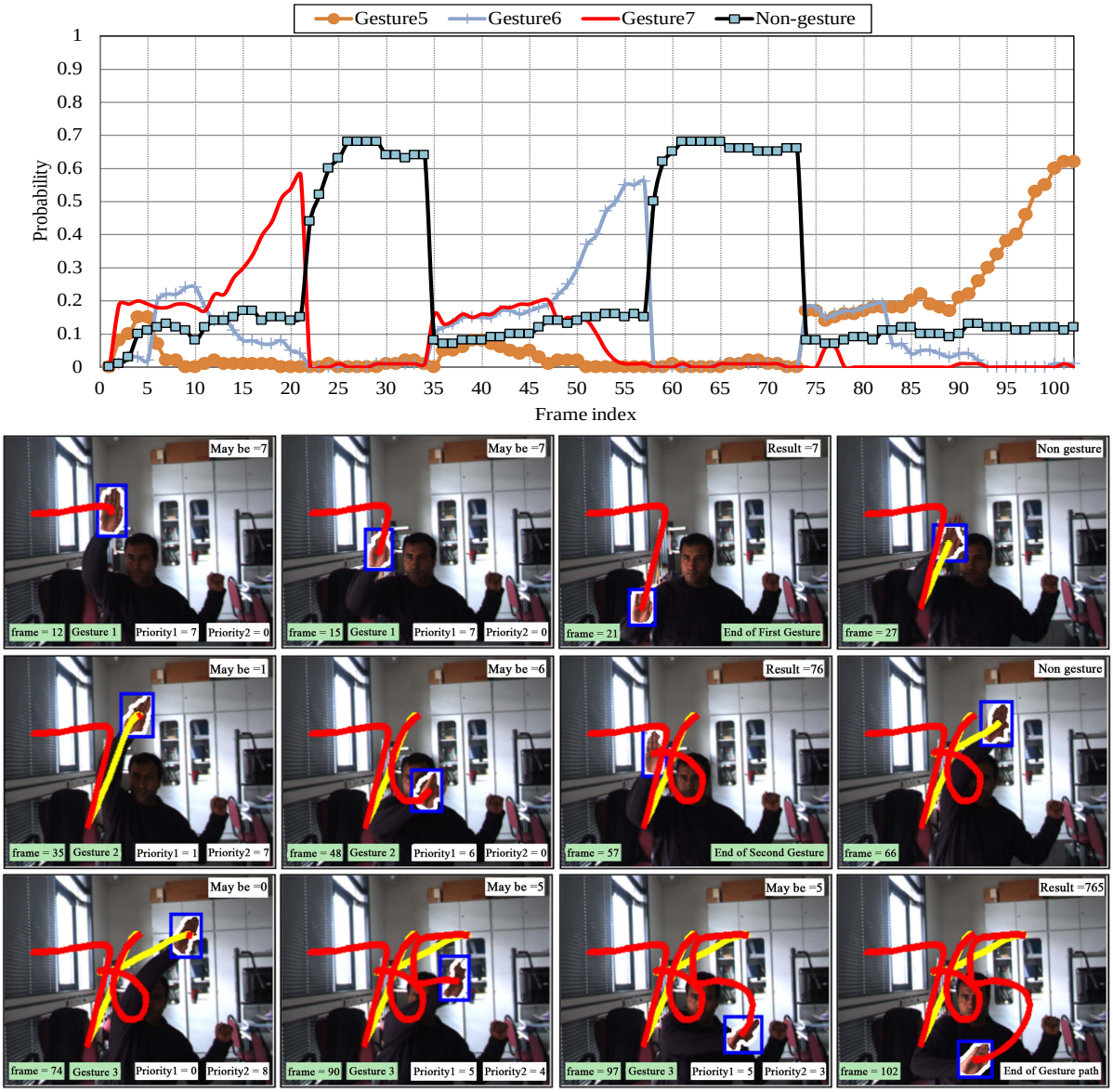


Figure 5. Temporal evolution of the probabilities of the gesture numbers “5”, “6”, “7” and non-gesture label “N”.

4. Conclusion

This paper proposed a novel gesture spotting and recognition technique, which handled hand gesture from continuous hand motion based on Conditional Random Field in conjunction with Support Vector Machine. 3D depth map was captured by bumblebee stereo camera to neutralize complex background sense. Additionally, dynamic affine-invariants features like elliptic Fourier and Zernike moments, in addition to three orientations motion features were employed to CRF and SVMs. Finally, the discriminative model of CRF performed the spotting and recognition processes by using the combined of orientation features. Accordingly, Support Vector Machine verified the hand shape at the start and the end point of meaningful gesture by using elliptic Fourier and Zernike moments features. Experiments showed that our proposed method could successfully spot hand gesture from continuous hand motion data with 92.50% recognition rate.

References

[1] Rose, R.C. (1992) Discriminant Word Spotting Techniques for Rejection Non-Vocabulary Utterances in Unconstrained

Speech. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **2**, 105-108.

[2] Chen, F.R., Wilcox, L.D. and Bloomberg, D.S. (1993) Word Spotting in Scanned Images Using Hidden Markov Models. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, **5**, 1-4. http://dx.doi.org/10.1109/icassp.1993.319732

[3] Starner, T., Weaver, J. and Pentland, A. (1998) Real-Time American Sign Language Recognition Using Desk and Wearable Computer Based Video. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, **20**, 1371-1375. http://dx.doi.org/10.1109/34.735811

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

[5] Baudel, T. and Beaudouin, M. (1993) CHARADE: Remote Control of Objects Using Free-Hand Gestures. *Communications of ACM*, **36**, 28-35. http://dx.doi.org/10.1145/159544.159562

[6] Wexelblat, A. (1994) Natural Gesture in Virtual Environments. *Proceedings of Virtual Reality Software and Technology Conference*, Singapore, 23-26 August 1994, 5-16. http://dx.doi.org/10.1142/9789814350938_0002

[7] Lee, H.-K. and Kim, J.H. (1999) An Hmm-Based Threshold Model Approach for Gesture Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **21**, 961-973. http://dx.doi.org/10.1109/34.799904

[8] Yang, H.-D., Sclaroff, S. and Lee, S.-W. (2009) Sign Language Spotting with a Threshold Model Based on Conditional Random Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **31**, 1264-1277. http://dx.doi.org/10.1109/TPAMI.2008.172

[9] Elmezain, M. (2013) Adaptive Foreground with Cast Shadow Segmentation Using Gaussian Mixture Models and Invariant Color Features. *International Journal of Engineering Science and Innovative Technology* (*IJESIT*), **2**, 438-445.

[10] Elmezain, M., Al-Hamadi, A., Niese, R. and Michaelis, B. (2009) A Robust Method for Hand Tracking Using Mean-Shift Algorithm and Kalman Filter in Stereo Color Image Sequences. *International Conference on Computer Vision, Image and Signal Processing, PWASET*, **59**, 355-359.

[11] Ding, C. and He, X.F. (2004) *K*-Means Clustering via Principal Component Analysis. *Proceedings of the* 21*st International Conference on Machine Learning*, New York, 225-232.

[12] Nixon, M.S. and Aguado, A.S. (2002) Feature Extraction and Image Processing. Newnes, Central Tablelands.

[13] Ahmad, M. and Lee, S.-W. (2008) Human Action Recognition Using Shape and Clg Motion Flow from Multi-View Image Sequences. *Journal of Pattern Recognition*, **41**, 2237-2252. http://dx.doi.org/10.1016/j.patcog.2007.12.008

[14] Dugad, R., Ratakonda, K. and Ahuja, N. (1998) Robust Video Shot Change Detection. *Workshop on Multimedia Signal Processing*, Redondo Beach, 7-9 December 1998, 376-381. http://dx.doi.org/10.1109/mmsp.1998.738965

[15] Dalal, N. and Triggs, B. (2005) Histograms of Oriented Gradients for Human Detection. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Diego, 25-25 June 2005, 886-893. http://dx.doi.org/10.1109/cvpr.2005.177

[16] Elmezain, M., Al-Hamadi, A. and Michaelis, B. (2010) Robust Methods for Hand Gesture Spotting and Recognition Using Hidden Markov Models and Conditional Random Fields. *IEEE Symposium on Signal Processing and Information Technology* (*ISSPIT*), Luxor, 15-18 December 2010, 131-136. http://dx.doi.org/10.1109/ISSPIT.2010.5711749