Hand Gesture Segmentation in Uncontrolled Environments with Partition Matrix and a Spotting Scheme based on Hidden Conditional Random Field

Yi Yao and Chang-Tsun Li
Department of Computer Science,
University of Warwick,
Coventry, UK, CV4 7AL
Y.Yao@warwick.ac.uk, c-t.li@warwick.ac.uk

***Abstract*—Hand gesture segmentation is the task of interpreting and spotting meaningful hand gestures from continuous hand gesture sequences with non-sign transitional hand movements. In real world scenarios, challenges from the unconstrained environments can largely affect the performance of gesture segmentation. In this paper, we propose a gesture spotting scheme which can detect and monitor all eligible hand candidates in the scene, and evaluate their movement trajectories with a novel method called Partition Matrix based on Hidden Conditional Random Fields. Our experimental results demonstrate that the proposed method can spot meaningful hand gestures from continuous gesture stream with 2-4 people randomly moving around in an uncontrolled background.**

Keywords—hand gesture spotting; uncontrolled environments; Hidden Conditional Random Fields

I. INTRODUCTION

Hand gesture recognition is one of the emerging and promising research fields of computer vision. For the continuous gesture streams, hand gesture recognition methods are not capable of detecting the starting and ending points of the meaningful gestures, while hand gesture segmentation (or hand gesture spotting) is the task of detecting and recognising hand gestures from signed utterances. In real world applications, the challenges from uncontrolled environments can largely affect the performance of hand gesture spotting methods, including: gesturing hand out of the scene, complex background, skin-coloured regions moving in background, performers wearing short sleeve and face overlapping with hand. In this paper, we propose a gesture spotting scheme especially for hand gesture spotting in uncontrolled environments.

Elmezain et al. [1] proposed a spotting scheme based on Conditional Random Fields (CRF). This method can only cope with perfectly controlled environments, and there are no latent variables to learn the transition probabilities between different segments of gestures. An adaptive threshold method utilising non-sign models based on CRF for hand gesture recognition and spotting was proposed in [2]. It is also only capable of handling controlled environments. Another gesture spotting method for uncontrolled environments was proposed in [3]. The method needs predefined number of hand candidates to deal with distractions in the background. The proposed method does not require any prior knowledge on the environments. Our experimental results show that the proposed method can segment and recognise meaningful hand gestures from severely distracted background.

II. TRACKING SCHEME

The tracking scheme of the proposed spotting method is able to detect and track all eligible hand candidates in the scene. In the first frame, Viola-Jones face detector [4] is used to detect eligible face regions. Then the pixels within the face regions are used to estimate the skin colour cues in the HSV colour space. Hence this tracking scheme can adapt to changing lighting conditions. Regions of Interests (ROIs) are then detected (Figure 1) as large exterior contours in skin colour binary image of the first frame. Then the Speeded Up Robust Features (SURF) [5] are extracted from adjacent frames to locate the same ROIs. The movement direction of ROIs are calculated as the only trajectory feature for the spotting scheme.

III. SPOTTING SCHEME

Since our tracking scheme is able to monitor all eligible hand candidates in the scene to tackle the challenge of moving objects in the background, a novel spotting scheme based on Hidden Conditional Random Fields (HCRF) [6] is proposed specifically for working with our tracking scheme and accomplishing hand gesture spotting in uncontrolled environments.

Before the gesture spotting process, the Non-sign model for HCRF is built with our novel method, and no training process is needed. From the first frame of the input video, the proposed spotting scheme takes one frame at a time, and uses sliding windows 'SWs' to extract video fragments with different lengths that start from the current frame of the input video stream. A weighting model, called Partition Matrix is then combined with Non-sign model to evaluate all video fragments. The gesture scores of all gesture classes at the current frame are then calculated based on scores of all previously extracted video fragments. The proposed spotting scheme will determine the starting and ending points of potential meaningful gestures with the scores on frames.

978-1-4799-2190-4/13 $31.00 © 2013 IEEE
DOI 10.1109/ACPR.2013.153

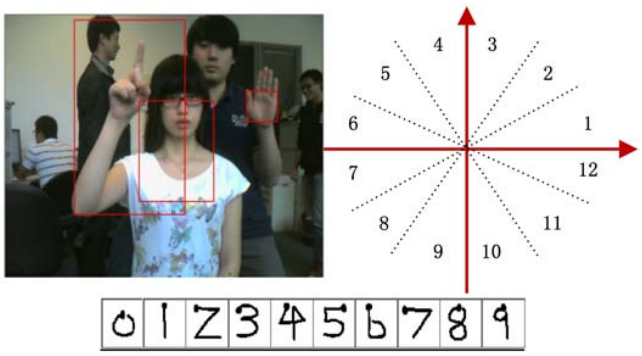

Fig. 1. Left: Initial ROIs; Right: The movement directions. Below: the definition of the gesture set in our experiments.

A. Partition Matrix with Non-Sign Model

HCRF models are undirected graph models that can learn the transition probabilities between different modules of the observation sequences, with hypothetical states. In our experiments, the task is recognising 10 hand-signed digits $Y=[y_0,y_1,\ldots,y_9]$, as shown in Figure 1, and the hidden states $H=[h_0,h_1,\ldots,h_S]$ are defined as the strokes of the gestures. We defined 13 hidden states from our gesture classes. The optimisation method for training the weight vector θ of the potential function is Limited Memory Broyden–Fletcher–Goldfarb–Shanno (Limited Memory BFGS) method [7]. The weight vector is initialised with the mean value and the regularisation factors are set to zero. In the original HCRF, the probability of gesture y given θ, h and the trajectory feature vector of a video fragment, $X=\{x_i \mid i=0,1,\ldots,N-1\}$, is:

$$P(y\mid X,\theta)=\sum_h P(y,h\mid X,\theta)=\frac{\sum_h \exp\{\Psi(y,h,X;\theta)\}}{\sum_{y',h}\exp\{\Psi(y,h,X;\theta)\}} \quad (1)$$

The partition function which can be understood as the score for y, given X and the trained HCRF model is defined as:

$$Z(y\mid X,\theta)=\sum_h \exp\{\Psi(y,h,X;\theta)\} \quad (2)$$

where the potential function $\Psi(y,h,X;\theta)$ is defined as:

$$\Psi(y,h,X;\theta)= \sum_j\left[\sum_k^{N_{XS}}\tau_k f_1(x_j,h_j)+\sum_m^{N_{SY}}\mu_m f_2(h_j,y)+\sum_g^{N_T}\lambda_g f_3(h_j,h_{j-1},y)\right] \quad (3)$$

where j is the frame number of X, E is the set of adjacent states, f_1, f_2 and f_3 are the three types of feature functions. Also, we assume the trained weights of feature functions have the form: $\theta=\{\tau_1,\tau_2,\ldots,\tau_{N_{XS}};\mu_1,\mu_2,\ldots,\mu_{N_{SY}};\lambda_1,\lambda_2,\ldots,\lambda_{N_T}\}$, where N_{XS}, N_{SY} and N_T are the total number of the three types of feature functions. f_1, f_2 and f_3 represent state features of compatibility of certain state with trajectory features, compatibility of certain state with gesture classes and transition features of certain combination of two states with gesture classes, respectively (for details please see [6]).

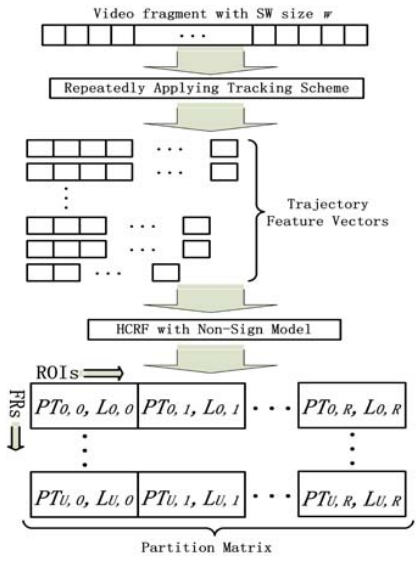

Fig. 2. Process of calculating partition values for a video fragment with SW size of w.

To segment meaningful gestures from transitional hand movements, we define a dedicated class label y_N for garbage gestures. Hence, the gesture class set becomes $Y=[y_0,y_1,\ldots,y_9,y_N]$. A method is used to build correspondent feature functions for this non-sign gesture class, which calculates the weights of the new feature functions based on trained weights of existing feature functions. The reasons for not including non-sign samples in the training stage for learning non-sign gesture patterns in HCRF are: firstly, it is nearly impossible to collect all meaningless hand movement patterns, due to the infinitely possibilities of random movements; secondly, since the task is hand gesture recognition in uncontrolled environments, the tracking results are distorted by background distractions on certain extent, which makes the states and transition features learned by HCRF distorted. Hence the most accurate way of learning the weights of non-sign feature functions is estimating them from the weights of meaningful gesture feature functions. In other words, learning non-sign gestures' features as the features that are not consistent with meaningful gestures.

Two series of new feature functions for non-sign gestures are used in the proposed method. One is a series of state feature $f_2(h,y_N)$, which represents the compatibility between hidden states and y_N. The weights of these new state features are calculated from trained weights of existing state features $f_2(h,y_{0-9})$:

$$\mu_m(h_j,y_N)=\bar{\mu}_m(h_j,y_{0-9})+T\cdot\sqrt{\sigma_m(h_j,y_{0-9})} \quad (4)$$

where $\bar{\mu}_m(h_j,y_{0-9})$ is the mean of weights of the state features $f_2(h_j,y_{0-9})$, which are features of state h_j and all 10 meaningful gesture classes. $\sigma_m(h_j,y_{0-9})$ indicates the variance of $f_2(h_j,y_{0-9})$. These weights indicate the significance of the appearance of certain state in the non-sign movements. The other series of

new feature functions is transition feature functions of adjacent states and y_N. The weights are calculated as:

$$\lambda_g(h_j,h_k,y_N)=\bar{\lambda}_g(h_j,h_k,y_{0-9})+T\cdot\sqrt{\sigma_g(h_j,h_k,y_{0-9})} \quad (5)$$

where $\bar{\lambda}_g(h_j,h_k,y_{0-9})$ and $\sigma_g(h_j,h_k,y_{0-9})$ are the mean and variance of weights of transition features $f_3(h_j,h_k,y_{0-9})$ respectively. The scale factor *T* is set to 1.2 empirically.

A weighting model called Partition Matrix is introduced to improve the performance of HCRF with distorted tracking results of multiple hand candidates in uncontrolled environments. As shown in Figure 2, the idea of Partition Matrix is to classify gestures with combined tracking results from different ROIs and frame selection patterns. It is obvious that the textures of the unconstrained background are changing rapidly with all the randomly moving objects. Since the texture features are used in our tracking scheme, applying tracking scheme with different frame selection patterns, namely different frame rates, will capture the texture of target hand with relatively stable changing rate. The texture of background distractions, on the other hand, have more intensely changing rates under different frame rates. Assuming that $V=\{f_i \mid i=0,1,2,\ldots,N-1\}$ is the current input video, f_n is the n^{th} frame. We define the video with frame rate FR_p is:

$$FR_p=\{f_i \mid i=0,p,2p,3p\ldots\} \quad (6)$$

which is a subset of *V*. Frame rates FR_1 to FR_4 are used to create different tracking results in our experiments. The tracking results $X=\{x_{u,r} \mid u=0,1,\ldots U, r=0,1,\ldots R\}$ is used to build the Partition Matrix, where $x_{u,r}$ is a trajectory feature vector with u^{th} frame rate and r^{th} hand candidate. In order to compare partition values from tracking results with different lengths, a normalisation function is introduced:

$$PT(y \mid x_{u,r},\theta)=\sum_h \exp\{\Psi(y,h,x_{u,r};\theta)\}/n_{u,r} \quad (7)$$

where $n_{u,r}$ is the amount of frame in $x_{u,r}$. Then we can introduce the definition of Partition Matrix:

$$\begin{pmatrix} (PT_{0,0},L_{0,0}) & \cdots & (PT_{0,R},L_{0,R}) \\ \vdots & \ddots & \vdots \\ (PT_{U,0},L_{U,0}) & \cdots & (PT_{U,R},L_{U,R}) \end{pmatrix} \quad (8)$$

where $(PT_{u,r},L_{u,r})$ is a partition-label pair of $x_{u,r}$. The purpose of using Partition Matrix and Non-sign Model is to calculate the partition values of all gesture classes for the video fragments. A new partition function is defined for this purpose:

$$Z'(y \mid X,\theta)=\sum_{x_{u,r}\in X}\{PT(y \mid x_{u,r},\theta)\cdot w_{u,r}\} \quad (9)$$

$w_{u,r}$ is the weight for the partition value, which is defined as:

$$w_{u,r}=1+W_{FR}(x_{u,r})+W_{ROI}(x_{u,r}) \quad (10)$$

$W_{FR}(x_{u,r})$ and $W_{ROI}(x_{u,r})$ are the Frame Rate Weight function and ROI Weight function:

$$W_{FR}(x_{u,r})=\begin{cases}1/U, x_{u,r}=\arg\max_{x_{u,r}\in\{x_{u',r}\mid u'=u\}} PT(y \mid x_{u,r},\theta) \\ 0, otherwise\end{cases} \quad (11)$$

$$W_{ROI}(x_{u,r})=\left|\{x_{u',r'} \mid W_{FR}(x_{u',r'})\neq 0, r'=r\}\right|/U \quad (12)$$

Frame Rate Weight function gives a larger weight to the trajectory feature vector with maximum $PT(y \mid x_{u,r},\theta)$ value within a row of the Partition Matrix. ROI Weight function represents the confidence of the r^{th} ROI being the target hand.

B. Multiple Sliding Windows Forward Spotting Scheme

A forward spotting scheme is proposed in this paper to determine the starting and ending frames of the meaningful gestures in continuous gesture video stream with uncontrolled environments. Figure 3 demonstrates the structure of the proposed spotting scheme.

As the new input video comes in, starting from the first frame, the proposed spotting scheme takes a series of sliding windows with different sizes to extract video fragments from the input video. The video fragment starts from frame f_c with the sliding window size *g* is defined as:

$$FSW_{c,g}=\{fi \mid i=c,c+1,c+2,\ldots,c+AL_g\} \quad (13)$$

where AL_g is the average length of all the training sample of the gesture *g*. Hence there are in total 10 different sliding window sizes. Unlike CRF, HCRF models do not produce gesture labels for every frame. For evaluating the probabilities of every frame being each gesture classes, the sizes of the sliding windows must cover the length of all gesture classes, instead of fixed size [1]. Since the HCRF models are trained from the training set, the meaningful gesture video fragments with similar amount of frames as the samples in the training set will have larger difference between the partition value of correct label and other labels. In other words, those video fragments can induce the best performance of the trained HCRF model. Also, the sliding window sizes preserve the ratio of average length of the gesture classes, and the partition values of the video fragments depend on their length (Equation 2,3). Therefore, the normalised partition values (Equation 7) of video fragment set $FSW_{c,0-9}$ can be seen as the scores of the current frame f_c being part of the meaningful gestures 0 to 9.

After video fragment set $FSW_{c,0-9}$ is extracted, Partition Matrix with Non-Sign Model is used to evaluate every video fragment. Then a matrix is formed, as every column is the normalised partition values of one video fragment against all gesture classes:

$$\begin{pmatrix} (PT_{y_0,0}) & \cdots & (PT_{y_0,9}) \\ \vdots & \ddots & \vdots \\ (PT_{y_N,0}) & \cdots & (PT_{y_N,9}) \end{pmatrix} \quad (14)$$

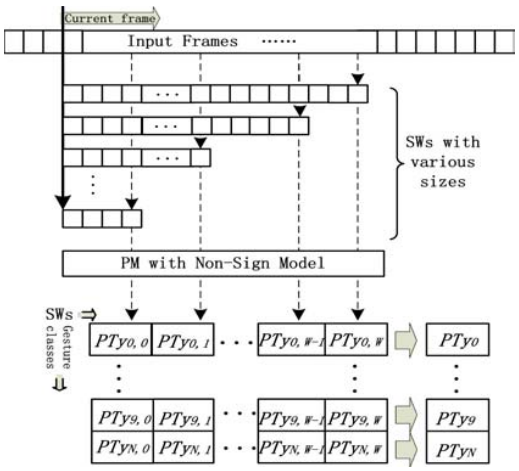

Fig. 3. Structure of the proposed spotting scheme.

The final partition values of all gesture classes PT_{y_0} to PT_{y_N} of the current frame f_c are calculated as:

$$PT_y = \sum_{w=0}^{|\gamma|} PT_{y,w} \tag{15}$$

namely, the sums of all elements in each row of the matrix (Equation 14). For the different sizes of the sliding windows, the partition value of the video fragment against certain gesture class would vary. When the size of sliding window tend to be more similar with the actual size of the meaningful gesture that the current frame belongs to, the partition value would increase. Hence, multiple video fragments with similar sizes of the actual target meaningful gesture, would have relatively higher partition values than video fragments with other sizes. Hence, for target meaningful gesture *y*, by adding up all partition values $PT_{y,0}$ to $PT_{y,W}$ of the current frame f_c, the final partition value PT_y is higher than the final partition value only using single sliding window size or average partition values of all sliding window sizes. Therefore, using the summation (Equation 15) increases the spotting performance.

As the tracking scheme combined with the Partition Matrix, the proposed spotting scheme is able to capture the target meaningful gesture trajectories from multiple hand candidates in uncontrolled environments, while the method of [1] is only capable of handling the existence of single hand candidate. Follow the work of [1], differential probability (DP) function is also used to determine the starting and ending frames of meaningful gestures. In this paper, we propose a new DP function. Instead of the normal inference results of the single video fragment, for all gesture classes, the summations of all normalised partition values of all video fragments with different sliding window sizes, namely PT_{y_0} to PT_{y_N}, are used to calculate the DP value:

$$DP(f_c) = \max_{y=0,1,..,9} PT_y - PT_{yN} \tag{16}$$

If the DP value is positive at certain frame:

$$\exists y : PT_y > PT_{yN} \tag{17}$$

then this frame is treated as a starting point. At a later frame, when the DP value returns to negative:

$$\forall y : PT_y < PT_{yN} \tag{18}$$

then this later frame is treated as an ending point. The gesture recognition result of the accumulative video fragment between the starting and ending point is calculated by Partition Matrix with the same procedure shown in Figure 2, but without Non-Sign Model (For details of the accumulative video fragment please see [1]).

IV. EXPERIMENTS AND DISCUSSIONS

For testing the proposed method, a database is collected, to provide the proposed spotting scheme a uncontrolled environments with severely distracted unconstrained background. The training set contains 6 gesture performers. For each gesture class, each gesture performer signs the gesture 6 times. Therefore, there are in total 360 training samples (the training samples are manually segmented). All the training samples are collected with perfectly controlled environments, with controlled lighting and unified single colour background, without any kind of distraction in the background. Two testing sets are collected, one 'easy' set with the same scene setting as the training set, and one 'hard' set with uncontrolled background. In the hard set, the background is normal office scene without any lighting control, and the performers wear short sleeve tops. There are 2-4 people constantly and randomly walking around in the background, and deliberately making meaningless hand movements beside the gesture performer. For both testing sets, each of 6 performers signs gesture 0-9 continuously in one video sample for 4 times. Hence there are 240 gesture samples in both easy and hard sets. The method of [1] is our inspiration, hence it is implemented for comparison in our experiments. The tracking results fed into our implementation of [1] are from our own tracking scheme, and the tracking results of the target hand ROI is manually picked out. That makes the experiments fair for the purpose of comparing spotting schemes.

Figure 4 shows how meaningful gesture can be detected from distracted background with other hands moving. Figure 4(a) illustrates the trajectory features of all samples of gesture 6 in the training set, while Figure 4(b) shows the tracking results of all hand candidates in a testing sample of hard set. The red line indicating the target ROI, and we can see from frame 4 to 63, a gesture 6 trajectory is detected.

The experimental results of the proposed spotting scheme on the hard and easy testing sets are shown in Table 1 and 2, and the comparisons with the performances of [1] on both easy and hard testing sets are shown in Table 3 and Figure 5. It is obvious that the proposed method outperformed [1].

V. CONCLUSIONS

In this paper, a novel gesture spotting scheme is proposed specifically for segmenting and recognising meaningful hand gestures in uncontrolled environments. A weighting model called Partition Matrix is used in conjunction with Non-Sign model to classify trajectories of all hand candidates. Experimental results show this spotting scheme can perform well in uncontrolled environments.

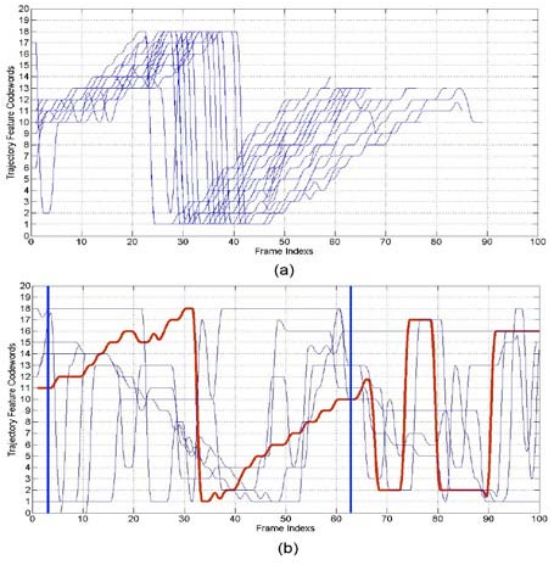

Fig. 4. (a): Trajectory features of all samples from gesture 6 in training set; (b): Tracking results of one sample in hard set, red line is target ROI.

TABLE I. RESULTS OF THE PROPOSED SPOTTING SCHEME ON HARD SET

Gesture class	Hard Set				
	Training samples	Testing samples	Recognition Results		
			True	Detected	Accuracy(%)
0	36	24	22	25	91.67
1	36	24	22	22	91.67
2	36	24	20	28	83.33
3	36	24	17	20	70.83
4	36	24	17	23	70.83
5	36	24	23	29	95.83
6	36	24	20	24	83.33
7	36	24	24	25	100.00
8	36	24	21	21	87.50
9	36	24	21	23	87.50
Overall	**360**	**240**	**207**	**240**	**86.25**

TABLE II. RESULTS OF THE PROPOSED SPOTTING SCHEME ON EASY SET

Gesture class	Easy Set				
	Training samples	Testing samples	Recognition Results		
			True	Detected	Accuracy(%)
0	36	24	23	23	95.83
1	36	24	22	22	91.67
2	36	24	23	24	95.83
3	36	24	22	22	91.67
4	36	24	21	26	87.50
5	36	24	22	24	91.67
6	36	24	21	25	87.50
7	36	24	24	26	100.00
8	36	24	24	25	100.00
9	36	24	23	23	95.83
Overall	**360**	**240**	**225**	**240**	**93.75**

TABLE III. COMPARISON OF PERFORMANCES

Warwick hand gesture database		
	Easy set	Hard set
Elmezain et al. ICPR 2010 [4]	92.08%	82.08%
The proposed method	**93.75%**	**86.25%**

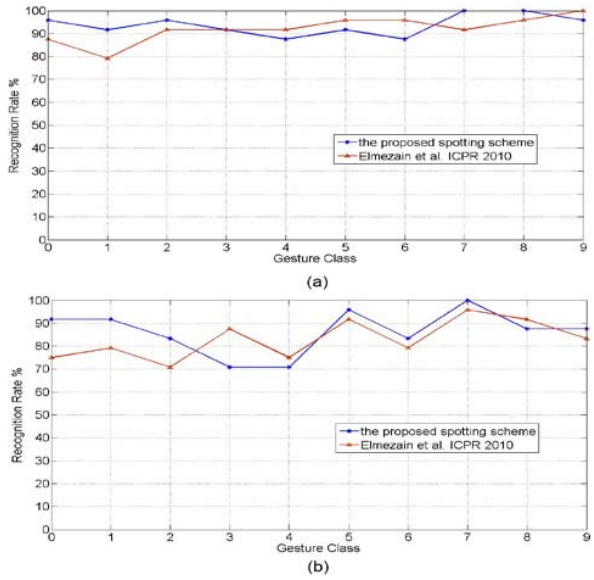

Fig. 5. Comparison of performances between the proposed spotting scheme and the method of [1]: Above: easy set; Below: hard set.

ACKNOWLEDGMENT

This work is included in the pending patent: UK patent application GB1305812.8, 28 March 2013, University of Warwick.

REFERENCES

[1] Elmezain, M., Al-Hamadi, A., Sadek, S., Michaelis, B, "Robust methods for hand gesture spotting and recognition using Hidden Markov Models and Conditional Random Fields", IEEE International Symposium on Signal Processing and Information Technology. ISSPIT' 2010.

[2] Hee-Deok Yang, Sclaroff, S., Seong-Whan Lee, "Sign Language Spotting with a Threshold Model Based on Conditional Random Fields", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, no.7, 1264-1277. 2009

[3] J. Alon, V. Athitsos, Q. Yuan, and S. Sclaroff, "A unified framework for gesture recognition and spatiotemporal gesture segmentation", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, no.9, 1685-1699. 2009

[4] Paul Viola, Michael J. Jones," Robust real-time face detection", International Journal of Computer Vision, Vol.57, 137-154. 2004.

[5] Herbert Bay, Andreas Ess, TinneTuytelaars, Luc Van Gool, "SURF: speeded-up robust features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3,346-359. 2008.

[6] A. Quattoni, S. Wang, L.P. Morency, M. Collins, and T. Darrell. "Hidden-state Conditional Random Fields", IEEE Transactions on Pattern Analysis and Machine Intelligence. vol.29, no.10, 1848 - 1852. 2007.

[7] Dong C. Liu, Jorge Nocedal," On the limited memory bfgs method for large scale optimization", Mathematical Programming, Springer-Verlag, Vol 45, Issue 1-3, 503-528. 1989.