Pattern Recognition 43 (2010) 197 – 203

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Action categorization with modified hidden conditional random field

Jianguo Zhang[a,*], Shaogang Gong[b]

[a]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, UK
[b]Department of Computer Science, Queen Mary University of London, London E1 4NS, UK

ARTICLE INFO

Article history:
Received 18 May 2007
Received in revised form 1 October 2008
Accepted 24 May 2009

Keywords:
Action recognition
Graph model
Hidden conditional random field
Optimum learning

ABSTRACT

In this paper, we present a method for action categorization with a modified hidden conditional random field (HCRF). Specifically, effective silhouette-based action features are extracted using motion moments and spectrum of chain code. We formulate a modified HCRF (mHCRF) to have a guaranteed global optimum in the modelling of the temporal action dependencies after the HMM pathing stage. Experimental results on action categorization using this model are compared favorably against several existing model-based methods including GMM, SVM, Logistic Regression, HMM, CRF and HCRF.

1. Introduction

Human action recognition is an important and challenging task. In general, there are two key elements in modelling human actions [13]: local appearance and temporal dependencies. To date, silhouette-based action recognition has been popular [13,23,20,3], i.e. an action is represented by a series of human body shapes. A silhouette is usually extracted by the estimation of the background, or given a known fixed background [3]. Other feature representations of action include space–time interest points [19,14], optical flow [7], motion template [4], and space–time volumes [25], shape context from still images [2,24], etc.

For learning the temporal dependencies between consecutive frames, numerous methods have been proposed with the vast majority based on graph models. Among them, hidden Markov model (HMM) is a baseline approach for modelling temporal dependencies. Its model parameters are estimated based on the optimization of the joint probability between the observations and sequence labels, which is marginalized over the hidden variables. Hence, it is a generative method, and not optimized based on the conditional Bayesian information. Though HMM has been shown good performance in many applications, for the purpose of pattern discrimination, an existing common consensus is that an ideal model should be derived and optimized based on maximizing the discrimination function [10]. Thus, to this point, HMM is not optimal. To overcome this limitation, conditional random field (CRF) was recently introduced [22,20]. However, CRF cannot incorporate the need for labelling a whole sequence as an action, and also cannot capture the intermediate structures using hidden state variables [15]. To overcome these problems, *hidden conditional random fields* (HCRF) was proposed by [11,24,15]. Compared to CRF, HCRF is capable of incorporating a sequence label into the optimization of observation conditional probabilities. However, due to the non-convexity nature of the objective function of HCRF, its performance depends heavily on its parameter initialization, thus not guaranteeing to give good results in a real application. To address these issues, in this work we formulate a modified HCRF (mHCRF) based on HMM pathing, and prove that the objective function of mHCRF is convex which a global optimum after the hidden variables become observable. This is the first contribution of our paper. We further develop an effective approach to silhouette-based action recognition using mHCRF. More specifically, we extract both a set of spectrum features using Fourier transform applied to the chain code of silhouettes and a set of motion moment features. The relationship between the whole sequence label and the temporal dependencies is then learnt using mHCRF. This is the second contribution of this paper. Thirdly, We compare this approach to other techniques including HMM, Gaussian mixture model, logistic regression, SVM, CRF and HCRF for action categorization.

2. Action features

Describing the silhouette of objects by their chain code has been widely adopted for shape retrieval/recognition or matching [9]. However, the chain code itself is not invariant to shape orientation change caused by possible 3D pose changes from human body actions. Specifically, if a human body shape is rotated by θ, the corresponding chain code $C(p)$ will be shifted by an offset, say,

* Corresponding author.
E-mail addresses: j.zhang@ecit.qub.ac.uk, jianguo.zhang@qub.ac.uk, jgzhang@dcs.qmul.ac.uk (J. Zhang), sgg@dcs.qmul.ac.uk (S. Gong).

0031-3203/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2009.05.015

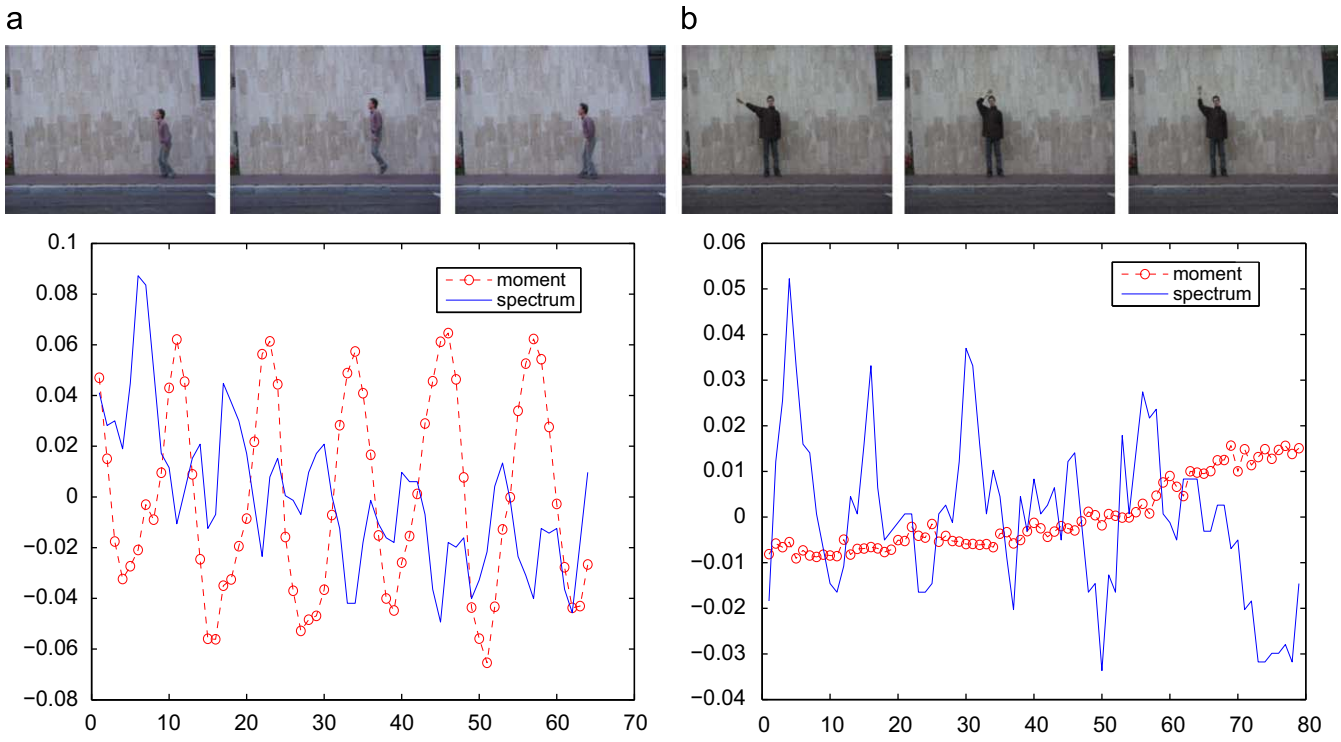

Fig. 1. Comparisons of action features extracted over time from: (a) a jumping sequence and (b) a waving sequence.

$\Delta p(\theta)$. Let $C(p+\Delta p(\theta))$ be the resulting chain code after rotation. To obtain rotation invariance, we perform as Fourier transform on $C(p)$, resulting $abs(F(C(p))) = abs(F(C(p+\Delta p(\theta))))$. The first n components of Fourier spectrum are selected as our action features, which we refer to as *spectrum* features in this paper.

Although these spectrum features are good for capturing actions that cause body shape change, e.g. in a bending or walking sequence, human body actions are not always necessarily associated with significant body shape change, e.g. jumping. In a jumping sequence, the human silhouette does not change a great deal over time resulting in its spectrum features less discriminative. To overcome this problem, we utilize motion moment features computed based on human silhouettes. Those silhouettes are extracted from an already known background. This simple procedure follows exactly the same way as [3]. Note there are more robust methods which can be used for silhouettes extraction of video sequences, e.g. the adaptive background detection [21]. Investigation of those methods beyonds the scope of this paper. Instead, we used the binary sequences directly available from [3].[1] We found that for binary images sequences, simple inter-frame differencing method performs well in our task. Here, for simplicity and computational proficiency, inter-frame differencing is used to detect motion changes resulting in a binary image, i.e. at time t, a binary image ΔB_t is obtained by the difference of two consecutive frames $\Delta B_t = B_{t+1} - B_t$. A set of moment features are then extracted from each binary image ΔB_t as follows:

$$A_t = \sum_{x,y} \Delta B_t[x,y], \quad \bar{x}_t = \frac{1}{A_t}\sum_{x,y} x\Delta B_t[x,y]$$

$$\bar{y}_t = \frac{1}{A_t}\sum_{x,y} y\Delta B_t[x,y], \quad E_t = \frac{\chi_{\max}}{\chi_{\min}} \tag{1}$$

[1] Those binary sequences are extracted from a known background in prior.

$$\chi^2 = \tfrac{1}{2}(a+c) + \tfrac{1}{2}(a-c)\cos 2\theta + \tfrac{1}{2}\sin 2\theta$$

$$a = \sum_{x,y}(x-\bar{x})^2 \Delta B_t[x,y]$$

$$b = \sum_{x,y}(x-\bar{x})(y-\bar{y})\Delta B_t[x,y]$$

$$c = \sum_{x,y}(y-\bar{y})^2 \Delta B_t[x,y]$$

$$\sin 2\theta = \pm\frac{b}{\sqrt{b^2+(a-c)^2}}$$

$$\cos 2\theta = \pm\frac{a-c}{\sqrt{b^2+(a-c)^2}} \tag{2}$$

$\chi_{\max}$ and $\chi_{\min}$ are selected among four possible values based on the sign of $\sin 2\theta$ and $\cos 2\theta$. To obtain shift invariance, we use $u_t = \bar{x}_{t+1} - \bar{x}_t$ and $v_t = \bar{y}_{t+1} - \bar{y}_t$. To make these moment features more robust to noise, we reject those features extracted from regions with small value A_t. A new value $A_t^* = A_{t-1}$ is associated to the rejected features and propagated from the previous frame. The rejection/propagation rule is defined as

$$A_t^* = \begin{cases} A_{t-1} & \text{if } A_t < tol \\ A_t & \text{otherwise} \end{cases} \tag{3}$$

In our experiments, we set the value of *tol* to 3. The corresponding values of other feature components are accordingly propagated. These moment features at time t are represented as $\{A_t^*, u_t*, v_t*, E_t^*\}$.

Fig. 1(a) shows an example where the moment features are more distinctive than the spectrum features in a jumping sequence, while Fig. 1(b) gives an example where the spectrum features are more distinctive than the moment features in a waving sequence. This indicates that these two types of features are complimentary. Our experimental results shown later further demonstrate this.

3. Hidden conditional random field

Hidden conditional random field was first introduced by Gunawardana et al. [11] for phone-conversation/speech classification and has then been applied to gesture and object recognition [24,15]. Given a sequence composed of a set of n local observations $\{x_1, x_2, x_3, \ldots, x_n\}$ denoted by $\mathbf{X}$, and its class labels $y \in Y$, we want to find a mapping $p(y|\mathbf{X})$ between them, where y is conditioned on $\mathbf{X}$. An HCRF is defined as

$$p(y|\mathbf{X};\Theta) = \frac{p(y,\mathbf{X};\Theta)}{p(\mathbf{X};\Theta)} = \frac{\sum_{\mathbf{H}} p(y,\mathbf{H},\mathbf{X};\Theta)}{\sum_{y,\mathbf{H}} p(y,\mathbf{H},\mathbf{X};\Theta)} = \frac{\sum_{\mathbf{H}} e^{\phi(y,\mathbf{H},\mathbf{X};\Theta)}}{\sum_{y,\mathbf{H}} e^{\phi(y,\mathbf{H},\mathbf{X};\Theta)}} \tag{4}$$

where Θ is the set of parameters of the model, and $\mathbf{H}=\{h_1, h_2, \ldots, h_n\}$. Each $h_i \in \hat{H}$ captures certain underlying structure of each class and $\hat{H}$ is the set of hidden states in the model. $\phi(y,\mathbf{H},\mathbf{X};\Theta)$ is the potential function which measures the compatibility between a label, a set of observations and a configuration of hidden variables. Based on maximum likelihood (ML) estimation, the regularized version of the objective function of HCRF (here, we minimize the equivalent negative log likelihood) is

$$\begin{aligned} L(\Theta) &= -\sum_{i=1}^{s} \log p(y_i|\mathbf{X}_i,\Theta) + \frac{\|\Theta\|^2}{2\sigma^2} \\ &= \underbrace{-\sum_{i=1}^{s} \log \sum_{\mathbf{H}} e^{\phi(y_i,\mathbf{H},\mathbf{X}_i;\Theta)}}_{(1)} \\ &\quad + \underbrace{\sum_{i=1}^{s} \log \sum_{y,\mathbf{H}} e^{\phi(y_i,\mathbf{H},\mathbf{X}_i;\Theta)}}_{(2)} + \underbrace{\frac{\|\Theta\|^2}{2\sigma^2}}_{(3)} \end{aligned} \tag{5}$$

where s is the total number of training sequences with known class labels. The first term and the second term are the log-likelihood of the data. The third term is the log of a Gaussian prior with variance σ^2, $p(\Theta) \sim \exp(\|\Theta\|^2/2\sigma^2)$, similar to regularization of a conditional random field [22]. The best parameters, $\Theta^* = \arg\min_\Theta L(\Theta)$, can be found by a gradient descent using Quasi-Newton optimization. Note that the objective function (Eq. (5)) and its gradient can be written in terms of marginal distributions over the hidden variables. These distributions can be computed exactly using inference methods such as belief propagation [22] when the graph model is a chain. From Eq. (5), we can see that the objective function of HCRF is not convex (non-negative sum of a concave function (term (1)) and two convex (terms (2) and (3)) functions does not guarantee convexity). Thus its global convergence heavily depends on the initialization. We shall address this problem in Section 5. Moreover, it is also very important to normalize each data first. Since the sum of potential could result in infinite values in the inference process for the gradient calculation, which could cause numerical instability.

4. Potential definition

In the context of action categorization, the potential function can be defined in terms of the following forms where observations interact with the hidden states and the sequence class labels interact with both the individual hidden node and the edges between hidden nodes:

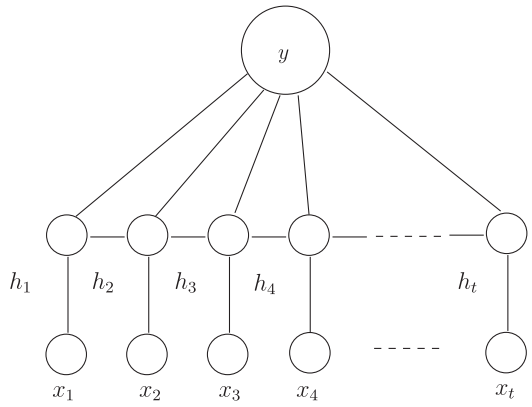

Fig. 2. The graph model of HCRF.

$$\phi(y,\mathbf{H},\mathbf{X};\theta) = \sum_j f(x_j)\cdot\theta(h_j) + \sum_j f(h_j)\cdot\theta(y,h_j) + \sum_{e_k\in E} f(e_k)\cdot\theta(y,e_k) \tag{6}$$

where e_k is an edge between a pair of nodes j and j'. In action recognition, the HCRF graph model is defined as a chain where each node corresponds to a hidden state variable at time t. $f(x_j)$ is a feature vector of node j. $f(h_j)$ is the feature vector corresponding to the hidden node j. $f(e_k)$ is the feature vector corresponding to the edge between node j and j'. Fig. 2 shows the graph model of HCRF as an undirected graph.

5. Modified HCRF

As stated in Section 3, due to non-convexity of the objective function (Eq. (5)), the initial parameters of HCRF must be carefully selected. This limits its usefulness. To overcome this problem, we seek an alternative approach. The idea is to make those hidden variables observable under the condition of learning HMM. Once the hidden variables become 'observable' to HCRF, the objective function can be shown to be convex. We describe our approach in two steps.

5.1. Automatic HMM pathing

First, we learn an HMM for each action class. The number of hidden states is automatically selected by a Gaussian mixture model using minimum description length (MDL) [17]. We then compute the Viterbi path for each training sequence. Here we refer to this step as *HMM pathing*. Thus the node of each training sequence is labelled by the learnt class specific HMM and this procedure makes the hidden states 'observable'. Our observed feature vector is continuous in $\mathbb{R}^L$, and we choose a Gaussian Mixture based HMM for the pathing stage. So a class specific HMM can be learnt by maximizing the following criteria:

$$\begin{aligned} \Theta^*(c) &= \arg\max_{\Theta} \sum_{i=1}^{s(c)} p(\mathbf{X}_i|y=c;\Theta) \\ &= \arg\max_{\Theta} \log \sum_{i=1}^{s(c)} \sum_{\mathbf{H}} p(\mathbf{X}_i,\mathbf{H}|y=c;\Theta) \end{aligned} \tag{7}$$

and the observation model is

$$p(X_t = x|h_t = i) = N(x; u_i, \Sigma_i) = \frac{1}{(2\pi)^{d/2}\|\Sigma_i\|^{1/2}} \times \exp\left(-\frac{1}{2}(x-u_i)'\Sigma_i(x-u_i)\right) \quad (8)$$

the Viterbi path is inferred by $h^*_{1:t} = \arg\max_{h_{1:t}} p(h_{1:t}|x_{1:t})$, which makes the hidden states of each training sequence observable so we can use directly in the next learning step.

5.2. Discriminative learning with global optimum

After the HMM pathing stage, the model function of HCRF in Eq. (4) becomes

$$p(y|\mathbf{X};\hat{\Theta}) = \frac{p(y,\mathbf{X};\hat{\Theta})}{p(\mathbf{X};\hat{\Theta})} \stackrel{def}{=} \frac{p(y,\mathbf{H}(\mathbf{X});\hat{\Theta})}{\sum_y p(y,\mathbf{H}(\mathbf{X});\hat{\Theta})} = \frac{e^{\phi(y,\mathbf{H}(\mathbf{X});\hat{\Theta})}}{\sum_y e^{\phi(y,\mathbf{H}(\mathbf{X});\hat{\Theta})}} \quad (9)$$

and the objective function (Eq. (5)) of HCRF becomes

$$L(\hat{\Theta}) = -\sum_{i=1}^{s}\log p(y_i|\mathbf{X}_i,\hat{\Theta}) + \frac{\|\hat{\Theta}\|^2}{2\sigma^2} = \underbrace{-\sum_{i=1}^{s}\phi(y,\mathbf{H}(\mathbf{X}_i);\hat{\Theta})}_{(L_1)} + \underbrace{\sum_{i=1}^{s}\log\sum_y e^{\phi(y,\mathbf{H}(\mathbf{X}_i);\hat{\Theta})}}_{(L_2)} + \underbrace{\frac{\|\hat{\Theta}\|^2}{2\sigma^2}}_{(L_3)} \quad (10)$$

We see that the effect of the HMM pathing stage is to change the relation between H, X in Eq. (4) into a function form in Eq. (9), that is the hidden variable becomes direct function of observation via $H(X)$. It is evident that the objective function $L(\hat{\Theta})$ is now convex.

Proof. From Eq. (10), we can have $L = L_1 + L_2 + L_3$.

- It is evident that L_1 is a linear function,when the potential function takes the form as in Eq. (11) and it can be viewed as a convex function.
- L_2 is a log–sum–exp term and it is a convex function.
- L_3 is quadratic function, and it is convex.
- Since the sum of the convex functions is convex, thus we have $L(\hat{\Theta}) = L_1 + L_2 + L_3$ convex.

Specifically, the first term is a linear function, can be viewed either concave or convex, the second and third term are convex. Because non-negative sum of two convex function guarantees convexity, $L(\hat{\Theta})$ becomes convex, which ensures that there is a global optimum. □

Accordingly, local potential of Eq. (6) is redefined as follows:

$$\phi(y,\mathbf{H},\mathbf{X};\hat{\Theta}) \stackrel{def}{=} \sum_j f(h_j)\cdot\theta_h(y,h_j) + \sum_{e_k} f(e_k)\cdot\theta_e(y,e_k) \quad (11)$$

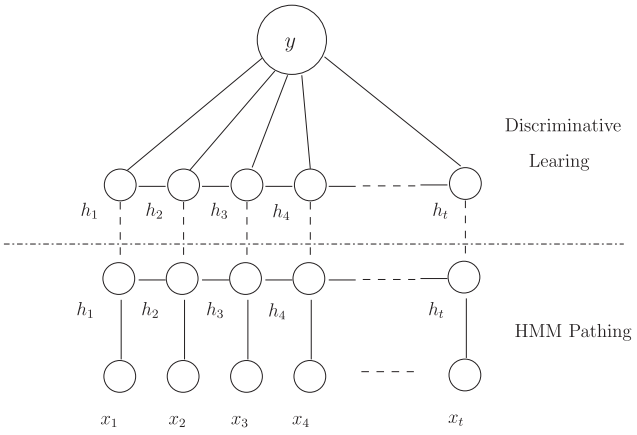

Fig. 3. The graph model of the proposed approach.

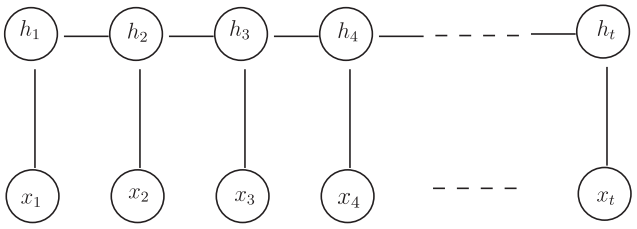

Fig. 4. The graph model of CRF.

Thus in our modified version of HCRF, parameter $\hat{\Theta}$ contains two components: $\hat{\Theta} = [\theta_h, \theta_e]$. We use $\theta_h[y, h_j]$ to refer to the parameter that measures the compatibility between a state h_j and an action label y. Similarly, $\theta_e[y, e_k]$ corresponds the parameter for compatibility between action label y and the edge between nodes j and j'.

We can see that the proposed approach still retains the advantage of a general HCRF, i.e. discriminative learning the parameters of the edge potentials (temporal dependencies) interacted with action sequence label. Moreover, Eq. (10) guarantees a global optimum. One may argue that the learning of HMM itself in the HMM pathing stage is not globally optimal. However, given the clear evidence that HMM has been successfully used in many applications, it could provide a better initialization for the convex objective function of mHCRF. The complexity of our approach is also reduced by making the hidden variables observable. It is obvious that for HCRF, when optimized using gradient descendant, the gradient has to be computed based on inference [24], whose complexity is usually exponential in the number of hidden variables. While in our approach, such inference is not necessary and the summation over the hidden variables is avoided.

The graph model of our proposed model is shown in Fig. 3. As a comparison, we also show the graph model of CRF in Fig. 4. Note that the objective function of CRF is $\log\sum_{i=1}^{s} p(\mathbf{H}|\mathbf{X}_i)$ ($\mathbf{H}$ is the observed node label instead of the whole sequence class label), which is different from ours as in Eq. (10).

Once we estimated the model parameters, the test of a new sequence is straightforward by maximizing the posterior probability of the learnt model with parameters $\hat{\Theta}^*$. Thus the final decision rule is

$$y^* = \arg\max_y p(y|\mathbf{X};\hat{\Theta}^*) \quad (12)$$

6. Experiments

We tested the effectiveness of the proposed method for action categorization. We compare the results from our model against those

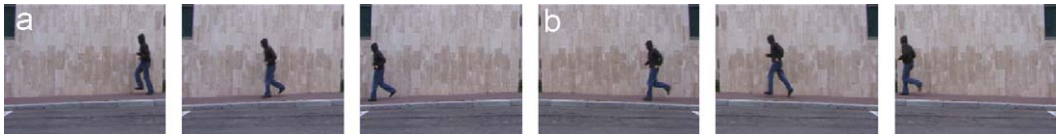

Fig. 5. Examples of misclassified sequences by the CRF approach: (a) a skip sequence and (b) a run sequence. All the two sequences are correctly classified by our approach.

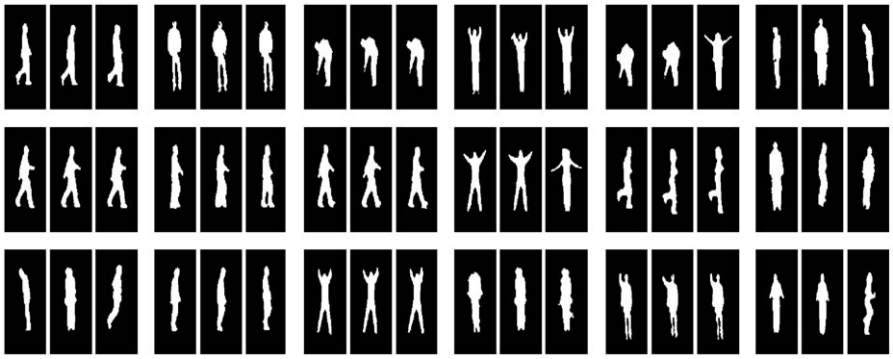

Fig. 6. Representative samples of each state in the HMM pathing stage, respectively.

of other existing action models including GMM [8], HMM [12,1], logistic regression (LR) [5], SVM [6] and CRF [20].

6.1. Data set

For this experiment, we use the data set from [3]. This data set contains 10 action classes with a total of 93 low resolution (180×144, 25 fps) video sequences showing nine different people, each performing 10 natural actions: 'running', 'walking', 'jumping-jack', 'jumping-forward-on-two-legs', 'jumping-in-place-on-two-legs', 'galloping-sideways', 'waving-two-hands', 'waving-one-hand', 'bending' and 'skipping'. We use all of the 10 action classes in our experiment.[2] Similar to [3], the silhouette of each frame is extracted based on subtraction of the median background from each of the sequences and a thresholding in color-space. The resulting silhouettes contained 'leaks' and 'intrusions' due to imperfect subtraction, shadows and color similarities with the background. The shape chain code is extracted for each silhouette, and Fourier transform is performed on the chain code data. The first 10 magnitudes of the Fourier response are used as the spectrum features. The motion moment features are extracted as described in Section 2. For the HMM pathing stage, we use the expectation maximization (EM) algorithm to optimize the model parameters. Fig. 6 shows some representative samples of each hidden state automatically discovered by HMM.

6.1.1. Action models

In our test, we compare several model-based approaches for action recognition. The models we compared include:

GMM: It is a generative model. We learn each class a Gaussian mixture model (GMM), i.e. $p(x;\Theta_c)$ for action class c. For a frame x_t at time t, we assign its class label based on the maximization of posterior probability, $c_t^* = \arg\max_c p(\Theta_c|x_t)$. For a sequence within time T, the class label of the whole sequence is determined by a majority voting strategy, $c^* = \arg\max_c(p(c))$ with $p(c) = (1/T)\sum_{t=1:T}\delta(c_t^* = c)$.

[2] This is different from the settings of [3], where they only used nine action classes.

In our experiments, the number of mixture components is automatically determined by using the MDL criteria. Note that GMM assumes the independency between local observations, thus it has no capability to model the temporal dependencies between consecutive frames.

Logistic regression (*LR*): Compared to GMM, Logistic regression is a simple but effective discriminative method in the family of graph models. Similar to GMM, it also assumes that there is no interaction between nodes. The difference to GMM is that it is optimized based on the conditional probability given its labels. The final class label of the whole sequence is determined in a similar way to that of GMM.

SVM: A widely used classifier and similar to logistic regression, it is a discriminative method without the consideration of the dependencies between frames, but optimized based on maximum separation margin between classes. We learn a multiclass SVM with RBF kernel, and then label each frame by the output of SVM. The optimal kernel parameters are found by cross validation over the training set. For a whole action sequence, the sequence label is determined in a similar way to that of GMM.

HMM: This model is capable of modelling the temporal dependence between hidden variables and it is a generative model. In our experiments, the number of hidden states and the transition matrix are automatically initialized using the MDL criteria over the whole training set. We then learn a HMM for each class respectively, denoted as M_c. Thus for a given sequence within time T, the class label is assigned via $c^* = \arg\max_c p(M_c|\mathbf{X})$.

CRF: Conditional random field (CRF) is a discriminative model with ability to learn the temporal dependencies between node labels. It is optimized based on the joint probability of node labels conditioned on the observations. We learn a single CRF for all the action classes, and then infer the Viterbi path for each test sequence. The label of the whole sequence is computed as the most frequently happened frame labels in the Viterbi path.

Table 3 shows the classification results by two fold cross validation on the data set. HMM gives better results than GMM due to the ability of modelling temporal dependencies. SVM performs better than LR since it is a non linear classifier while LR is a linear classifier in our case. Both LR and SVM perform better than GMM. Note the discriminative methods such as SVM and LR perform better than

Table 1
Confusion matrix of action classification results with CRF.

Action	Bend	Jack	Jump	Pjump	Run	Side	Walk	Wave1	Wave2	Skip
Bend	9									
Jack		9								
Jump			7							2
Pjump				7		1				1
Run		1			9					
Side						9				
Walk						1	9			
Wave1		1		1	1			4		2
Wave2		1			1				7	
Skip			3		1					6

The term 'jack' represents 'jumping-jack', 'pjump' for 'jumping-in-place-on-two-legs', 'side' for 'galloping-sideways', 'wave1' for 'waving-one-hand' and 'wave2' for 'waving-two-hands'.

Table 2
Confusion matrix of action classification results with mHCRF.

Action	Bend	Jack	Jump	Pjump	Run	Side	Walk	Wave1	Wave2	Skip
Bend	9									
Jack		8							1	
Jump			9							
Pjump		1		8						
Run					10					
Side						9				
Walk							10			
Wave1		1		1				5	2	
Wave2		1							8	
Skip			3							7

The term notations are the same as in Table 1.

Table 3
Classification accuracy of action categories with different methods and different features.

	Spectrum	Moments	Combination
GMM	0.628	0.651	0.604
HMM	0.791	0.605	0.744
LR	0.628	0.767	0.814
SVM	0.721	0.698	0.767
CRF	0.721	0.791	0.850
HCRF	0.781	0.847	0.880
mHCRF	0.800	0.865	**0.893**

HMM even though with the latter modelling temporal dependencies. CRF gives slightly better results than SVM. HCRF outperforms CRF due to its ability of discriminative learning of hidden states structures. Among all of the models, our method mHCRF performs the best. As to the features, we can see that the performance of moment features is usually better or comparable (in the case of SVM) than the spectrum features indicating that moment features have greater discrimination potential than spectrum features, with an exception in the case of HMM. This maybe caused by the over-fitting problem when learning HMM. The combination of the moment features with the spectrum features gives better results than using them alone. One point needs to be clear that the HMM 'pathing' stage in the mHCRF approach only outputs the optimal path (i.e. the optimal hidden states sequence) of the observation sequence based on Viterbi inference. It does not output the action label of the whole sequence. This is the main difference between HMM 'pathing' and HMM classifier used in the experiments. The common part is that training using the same MDL criteria. Thus the superior performance of mHCRF over HMM shown in Table 3 can also be considered as the additional gain of our approach with respect to the HMM pathing stage. Tables 1 and 2 shows CRF with all features and the confusion matrix of the results of using mHCRF, respectively. We can see the major errors are caused by the wave1 action and skipping action. Fig. 5 shows a run sequence and a skip sequence that are misclassified by the CRF approach. However, they are correctly classified by our approach despite of their similar visual appearance. Note that the results reported in [3] are not directly comparable to us. One reason is that in their paper, they use leave-one-out procedure which is much easier than two fold cross validation used in our experiments. Another reason is that we increased the difficulty of the task by including additional skipping sequence which is a more difficult class shown by the confusion matrix in Table 1. Using similar settings as in [3], we obtained the classification error rate of 1.1%, which is comparable to the results in [3] using advanced 3D space–time shape features.

7. Discussions and conclusions

In this work, we presented an action recognition model using a modified HCRF for a guaranteed global optimal solution after the HMM pathing stage. We showed that our model is effective and performs well against other existing techniques for action categorization.

In this paper, we focus on classification of action images as a whole, rather than identifying the detailed body configurations. It is worth noting that another line of research in motion action recognition is based on human parts [16]. In these approaches, the task of motion recognition can also be performed by firstly identifying the body configurations and then inferring the relationships between candidate body parts.

Our current approach focuses on the learning algorithm of action recognition. The robust silhouettes are extracted from a directly available background in the same way as in [3]. Note that there are some factors needed to be taken into account in the silhouettes extraction process when the background is unknown, e.g., shadow, illumination changes, camouflage. There is plenty of work done regarding those aspects in the silhouettes extraction process, e.g., long-term illumination changes can be modelled as a separate component

in the adaptive background modelling [21]. Furthermore, shadow can be removed by the method proposed by [18]. By addressing those issues, Zhuang et al. [26] further propose a method that could efficiently extract robust silhouettes of humans.

We still believe that developing good training algorithms for traditional HCRF will be another valuable direction. However, in this paper, we presented an alternative formulation of HCRF which at least have convex objective function at the second stage. The good initialization can be provided by the HMM pathing in practical. As to the action features, note that beside the motion moment features and spectrum features, there exists other features, e.g., shape context features [2]. Investigating their performance in the context of our framework will be an interesting direction. It is worth pointing out that actions could be better recognized across different scales, developing a multi-scale mHCRF as an extension of the proposed method is our future work.

References

[1] M. Ahmad, S.-W. Lee, HMM-based human action recognition using multiview image sequences, in: International Conference on Pattern Recognition, 2006, pp. 263–266.

[2] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (4) (2002) 509–522.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, in: International Conference on Computer Vision, 2005, pp. 1395–1402, URL: ⟨www.wisdom.weizmann.ac.il/vision/SpaceTimeActions.html⟩.

[4] A.F. Bobick, J.W. Davis, The recognition of human movement using temporal templates, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (3) (2001) 257–267.

[5] J.R. Brzezinski, Logistic regression modeling for context-based classification, in: DEXA '99: Proceedings of the 10th International Workshop on Database & Expert Systems Applications, 1999, p. 755.

[6] O. Chapelle, P. Haffner, V. Vapnik, Support vector machines for histogram-based image classification, IEEE Transactions on Neural Networks 10 (5) (1999) 1055–1064.

[7] A.A. Efros, A.C. Berg, G. Mori, J. Malik, Recognizing action at a distance, in: International Conference on Computer Vision, Nice, France, 2003, pp. 726–733.

[8] M.A.T. Figueiredo, A.K. Jain, Unsupervised learning of finite mixture models, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (3) (2002) 381–396.

[9] H. Freeman, On encoding arbitrary geometric configurations, IRE Transactions on Electronic Computers 10 (2) (1961) 260–268.

[10] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press Professional, San Diego, CA, USA, 1990.

[11] A. Gunawardana, M. Mahajan, A. Acero, J.C. Platt, Hidden conditional random fields for phone classification, in: International Conference on Speech Communication and Technology, 2005.

[12] H.-K. Lee, J.H. Kim, An HMM-based threshold model approach for gesture recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 21 (10) (1999).

[13] T.B. Moeslunda, A. Hiltonb, V. Krügerc, A survey of advances in vision-based human motion capture and analysis, Computer Vision and Image Understanding 104 (2–3) (2006) 90–126.

[14] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial-temporal words, in: British Machine Vision Conference, 2006.

[15] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, T. Darrell, Hidden-state conditional random fields, Technical Report, MIT, 2006.

[16] D. Ramanan, D.A. Forsyth, A. Zisserman, Tracking people by learning their appearance, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (1) (2007) 65–81.

[17] J. Rissanen, A universal prior for integers and estimation by minimum description length, Annals of Statistics 11 (2) (1983) 417–431.

[18] T. Matsuyama, S. Nobuhara, Y. Tsuda, I. Ohama, Multi-viewpoint silhouette extraction with 3d context-aware error detection, correction, and shadow suppression, in: Fourth European Conference on Visual Media Production (CVMP 2007), vols. 27 and 28, 2007, pp. 1–9.

[19] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local SVM approach, in: ICPR, Cambridge, UK, 2004, pp. 32–36.

[20] C. Sminchisescu, A. Kanaujia, D. Metaxasb, Conditional models for contextual human motion recognition, Computer Vision and Image Understanding 104 (2–3) (2006) 210–220.

[21] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real time tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, 1999.

[22] C. Sutton, A. McCallum, An introduction to conditional random fields for relational learning, in: L. Getoor, B. Taskar (Eds.), Introduction to Statistical Relational Learning, MIT Press, Cambridge, 2006, pp. 93–128.

[23] L. Wang, T. Tan, H. Ning, W. Hu, Silhouette analysis-based gait recognition for human identification, IEEE Transactions on Pattern Analysis and Machine Intelligence 25 (12) (2003).

[24] S.B. Wang, A. Quattoni, L.-P. Morency, D. Demirdjian, T. Darrell, Hidden conditional random fields for gesture recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1521–1527.

[25] A. Yilmaz, M. Shah, Actions sketch: a novel action representation, in: IEEE Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 2005, pp. 984–989.

[26] Y. Zhuang, C. Chen, Efficient silhouette extraction with dynamic viewpoint, in: International Conference on Computer Vision, 2007, pp. 1–8.

About the Author—JIANGUO ZHANG received his Ph.D. degree in 2002 from the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, PR China. He has been with the Department of Computer Science, Queen Mary University of London, UK (2005–2007), LEAR Group of INRIA Rhône-Alpes, France (2003–2005), the School of Electrical and Electronic Engineering, Nanyang Technological University of Singapore (2002–2003). He is currently a Lecturer of Visual Computing at Queen's University Belfast, UK. His research interests include pattern recognition, computer vision, visual surveillance, image processing and machine learning. He won the Best Paper Award (2008) of the International Machine Vision and Image Processing Conference.

About the Author—SHAOGANG GONG is Professor of Visual Computation at Queen Mary, University of London, elected a Fellow of the Institution of Electrical Engineers and a Member of the UK Computing Research Committee. He received his D.Phil. in 1989 from Keble College, Oxford University with a thesis on the computation of optic flow using second-order geometric analysis. He was a recipient of a Queen's Research Scientist Award in 1987, a Royal Society Research Fellow in 1987 and 1988, and a GEC-Oxford Fellow in 1989. He twice won the Best Science Prize of the British Machine Vision Conferences (1999, 2001), won the Best Paper Award (2001) of the IEEE International Workshop on Recognition, Analysis and Tracking of Faces and Gestures, and the Best Paper Award (2005) of the IEE International Symposium on Imaging for Crime Detection and Prevention. He has published over 180 papers in computer vision and machine learning, and a book on Dynamic Vision: From Images to Face Recognition. His work focuses on the detection, tracking and recognition of motion objects; video based face and expression recognition; gesture recognition for visually mediated interaction, video behavior profiling, recognition and abnormality detection.