Computer Vision and Image Understanding xxx (2010) xxx–xxx

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Action categorization by structural probabilistic latent semantic analysis

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

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

ARTICLE INFO

Article history:
Received 4 July 2007
Accepted 28 April 2010
Available online xxxx

Keywords:
Action categorization
pLSA
Structural pLSA
Local shape context

ABSTRACT

Temporal dependency is a very important cue for modeling human actions. However, approaches using latent topics models, e.g., probabilistic latent semantic analysis (pLSA), employ the bag of words assumption therefore word dependencies are usually ignored. In this work, we propose a new approach *structural* pLSA (SpLSA) to model explicitly word orders by introducing latent variables. More specifically, we develop an action categorization approach that learns action representations as the distribution of latent topics in an unsupervised way, where each action frame is characterized by a codebook representation of local shape context. The effectiveness of this approach is evaluated using both the WEIZMANN dataset and the MIT dataset. Results show that the proposed approach outperforms the standard pLSA. Additionally, our approach is compared favorably with six existing models including GMM, logistic regression, HMM, SVM, CRF, and HCRF given the same feature representation. These comparative results show that our approach achieves higher categorization accuracy than the five existing models and is comparable to the state-of-the-art hidden conditional random field based model using the same feature set.

1. Introduction

Human action recognition is an important and challenging task. Numerous methods have been proposed either focusing on building robust action representations or developing recognition models. In general, there are two key elements in modeling and recognizing human actions [24]: local appearance and temporal dependencies. In this section, we briefly review existing work in the area.

1.1. Motion representations

A great deal of work has been done to represent action sequences. They can be broadly classified into two categories: *part-based representations* and *holistic representations*. Part-based motion representation relates to the success of representing images using a sparse set of interesting points in object images [23], which offers invariance to rotational and affine changes, as well as robustness to background clutter. Laptev [19] proposed to interpret local motion changes by the neighborhoods of the spatial–temporal interest points. This work has been further extended to motion recognition by a bag of spatial–temporal words [25,30,20]. Laptev et al. [20] formulated the problem of motion recognition as a matching of corresponding events in image sequences based on an assumption that similar patterns of motion contain similar events with consistent motion across image sequences. The motion descriptor is constructed as a bag of local events. The local velocity adaptation of events allows this approach to recognize motion independently of the scale and Galilean transformations caused by the relative motion of the camera. One limitation of this approach is lack of modeling the relative structure of events in space time because all of the events are treated independently [20]. Dollar et al. [8] used a denser representation by sampling the interesting points as local maxima in the spatial direction only, and achieved better performance than a sparser representation. Using similar features, Niebles et al. [25] achieved comparable performance with an unsupervised approach, and Nowozin et al. [26] further improved the performance using a discriminative approach. It is worth noting that recent study has unveiled some limitations of interest points based approach for object recognition. When the object is small, the detectors cannot produce sufficient detections [10,42]. The detector of spatial–temporal interest point may suffer from insufficient detections as well when the motion scale is relatively small at a distance [9]. Therefore the performance of this approach under such circumstances is still not known.

Holistic approaches usually represent the motion sequences as a whole. These approaches often rely on dense or global representations. For global representations, a simple approach is to accumulate all measurements in a video sequence using a global descriptor, e.g., the motion history image (MHI) [3]. In the MHI based methods, the temporal influence of motion is encoded as the intensity differences in the MHI template where the recent

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

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.04.006

motions have higher intensity than the old ones. Inspired by this approach, Weinland et al. [40] developed a 3D motion representation called motion history volume using constrained multiple cameras. Boiman and Irani [4] proposed a motion detection approach by computing the correlation between two spatial–temporal volumes in a dense mode with a 3D sliding window approach. Efros et al. [9] used smoothed optical flow as the motion descriptor to classify human actions at a distance. Holistic approaches are known to be sensitive to the large geometrical variations between intra-class samples, moving cameras and non-stationary background. To handle these challenges, those approaches usually deploy some motion based segmentation techniques and then compute the motion descriptors in the segmented regions [9].

Another type of holistic approach is to represent actions as a sequence of human body shapes. This representation is inspired by the observation that a time series of 2D silhouettes in the space–time volume contains both the spatial configuration about the pose of human figure at anytime (location and orientation of the arms, legs, and torso, aspect ratio of different body parts) and the dynamic information (global body motion and motion of the limbs relative to body) [15]. To date, silhouette-based action recognition has received increasing attention [24,36,33,2]. For modeling objects, shape features contain rich and useful information of the object and have been demonstrated one of the key object descriptors for object detection in complex scenes [13,27]. Working in shape is also advantageous to other image representations, such as the appearance features [5]. For instance, Bosch et al. [5] have successfully incorporated a shape descriptor with the appearance features in a pyramid representation, which achieves the state-of-art recognition performance on the Caltech101 [11] and Caltech 256 [16] datasets with accuracies of 98.2% and 69.8% respectively. The object shape variations along time contain dynamic information of motion as well. Wang et al. [39] have demonstrated that it is promising to discover human motion categories from static images, whose features are represented as shape context [39]. Recently, Gorelick et al. [15] worked on 2D silhouette-based action sequences and represented actions as spatial–temporal shapes. Their approach exploited the solution to the Poisson shape representation to extract various shape properties for classification. Our motion representation is essentially a temporal sequence containing body shapes. The shape descriptor is inspired by the pioneering work of the shape context by Belongie et al. [1] and the success of bag of features representation for object recognition [10,42], thus resulting in a codebook representation of local context for each silhouette.

1.2. Motion models

Another important issue for motion recognition is to develop models to learn the temporal dependencies between consecutive frames. To date, numerous methods have been proposed with the vast majority based on graph models. Among them, the hidden Markov model (HMM) is a baseline approach for modeling temporal dependencies mainly interpreted by a transition matrix. Its model parameters are estimated based on the optimization of the class conditional joint probability distribution between the observations and the sequence labels, which is marginalized over a set of hidden variables. Hence, it is a generative method and not optimized based on the conditional Bayesian information. Though HMM has shown good performance in many applications, for the purpose of pattern discrimination, an existing common consensus is that an ideal model in theory should be derived and optimized based on maximizing the discrimination function [14]. Thus, to this point, HMM is not optimal. To overcome this limitation, conditional random field (CRF) was introduced recently [35,33]. However, CRF cannot incorporate the need for labeling a whole sequence as an action, and also cannot capture the intermediate structures using hidden state variables [29]. To overcome these shortcomings, *hidden conditional random field* (HCRF) was proposed in [17,38,29]. Compared to CRF, HCRF is capable of incorporating a sequence label into the optimization of the probabilities of sequence labels conditioned on observations. Recently, Nowozin et al. [26] proposed a discriminative subsequences mining approach to find the optimal discriminative subsequence patterns. In their approach, each video is encoded as a sequence of set of integers. To train a classifier on such a representation, they extended the PrefixSpan [28] subsequence mining algorithm in combination with LPBoost [7].

However, all of the above model-based approaches are *supervised*, i.e., the training set has to be manually labeled with varying degree of supervision. Thus, another interesting direction is to develop *unsupervised* learning methods, e.g., action recognition by *probabilistic latent semantic analysis* (pLSA) [25].

As one of the generic models, pLSA [31] has been successfully used to discover object categories without prior segmentation. For instance, Sivic et al. [31] used pLSA to automatically find the object categories from a large image collection. Fergus et al. [12] developed a translation and scale invariant pLSA model (TSI-pLSA) to localize an object from just its name, which extends pLSA to include spatial information. Inspired by the success of pLSA for object recognition, researchers have recently applied pLSA for motion classification. Niebles et al. [25] developed an unsupervised motion classification approach using pLSA with spatial–temporal words. However, the approach by Niebles et al. is not capable of localizing motion in videos. In order to tackle this problem and inspired by the successful Implicit Shape Model [22] for object detection, Wong et al. [41] recently extended the pLSA approach to localize motion categories by including the geometrical constrains between spatial–temporal patches, which is called pLSA-ISM. This approach is a promising extension of TSI-pLSA to infer the location of motion in video sequences. Experimental results show good localization performance with little presence of background clutter. However, the performance of this method in the presence of strong background motion clutter is still not known. For the purpose of statistical language modeling, Wang et al. [37] presented a directed Markov random field that combines *n*-gram models, probabilistic context free grammars and pLSA to learn the semantic information. However, as other MRF based approaches, the training process of this model needs the labels of language sequences.

The dynamic adjacent dependencies cannot be learnt explicitly by most of the pLSA based approaches, since the pLSA ignores such global dependencies in principle. To incorporate those dependencies into the unsupervised learning process by pLSA, we proposed a *structural pLSA* (SpLSA),[1] where we show that the standard pLSA is a special case of our model. We then develop an action categorization approach learnt by SpLSA with a codebook representation of local shape context. We compared our model with six other existing models using two standard action recognition datasets. In our experiments, all models were given exactly the same feature sets to remove any effect from the choice of different features.

The paper is organized as follows. Section 2 describes the principles of pLSA. We then present our new model SpLSA and the learning method in Section 3. Section 4 develops a novel action categorization approach based on SpLSA with a signature of codebook of local shape context. We evaluate our approach in Section 5. Section 6 concludes the whole paper and points out some future work.

[1] By 'structural', we mean conditional dependencies as in structural learning for probabilistic graph models.

Please cite this article in press as: J. Zhang, S. Gong, Action categorization by structural probabilistic latent semantic analysis, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.04.006

2. Probabilistic latent semantic analysis

Probabilistic latent semantic analysis (pLSA) was proposed in [18] and has been extensively studied as a model of a text document set. In this model, each document is generatively modeled as a bag of words, each of which is sampled from a document-specific mixture of Z latent 'topic' distributions. Each topic z is described by its distribution $p(w|z)$ over the W possible words of the dictionary and each document d is characterized by the mixture over Z topics. The T word instances of a document d are treated as a set of independent samples from the mixture. Letting z denote the unknown topic (mixture of component) of word w_i, the joint probability of the T words and corresponding d is modeled as follows:

$$p(w_1, w_2, \ldots, w_T, d) = \prod_{i=1}^{T} p(w_i|d)p(d) = \prod_{i=1}^{T} \sum_{z\in Z} p(w_i|z)p(z|d)p(d), \quad (1)$$

where Z is the set of possible topics. Besides the document categorization, pLSA has also been applied to computer vision, e.g., object recognition [32,12]. In those applications, the object images are considered as a mixture of topics and local patches (often produced by some interest point detectors [23]) are viewed as words, usually called visual words. Thus object images are modeled as a mixture of latent topics that generates each patch independently. Note that this model ignores the correlation between words based on the bag of words assumption.

The mixture density of the model, $p(w|z)$ and $p(z|d)$, can be learnt using an expectation maximization (EM) algorithm. The E step computes the posterior over the topic, $p(z|w,d)$ and then the M step updates the densities. Given a set of D documents, the model parameters are computed by maximizing the log of the following data likelihood of those documents:

$$L = \prod_{m=1}^{D} p(w_1, w_2, \ldots, w_{T_m}, d_m) = \prod_{m=1}^{D} \prod_{i=1}^{T_m} \sum_{z\in Z} p(w_i|z)p(z|d_m)p(d_m), \quad (2)$$

where T_m is the number of words of document d_m. Determining the most probable topic for a query image/document d is achieved by locking $p(w|z)$ and iterating with EM to estimate the $p(z|d)$.

3. Structural pLSA

As mentioned before, though, pLSA has achieved promising results in the application of automatic discovery of object categories, the original model ignores the relationship between words, i.e., it considers each word-document draw independent as in Eq. (1). However in real cases, this is somewhat limited, e.g., local patches of an object in a static image can have object-specific spatial constrains. For action recognition, it is evident that there exist strong connections between action words, i.e., the action clip at time t is highly related to other action clips occurred within the same short period. Thus in order to model such relationships in an unsupervised manner, we propose a *structural* pLSA (SpLSA). The idea is to incorporate the associations between words when modeling the documents, i.e., given a document d and its corresponding words $w_1, w_2, \ldots, w_T$, the joint probability is described as follows when the probabilistic graph represents a Bayesian network as shown in Fig. 1:

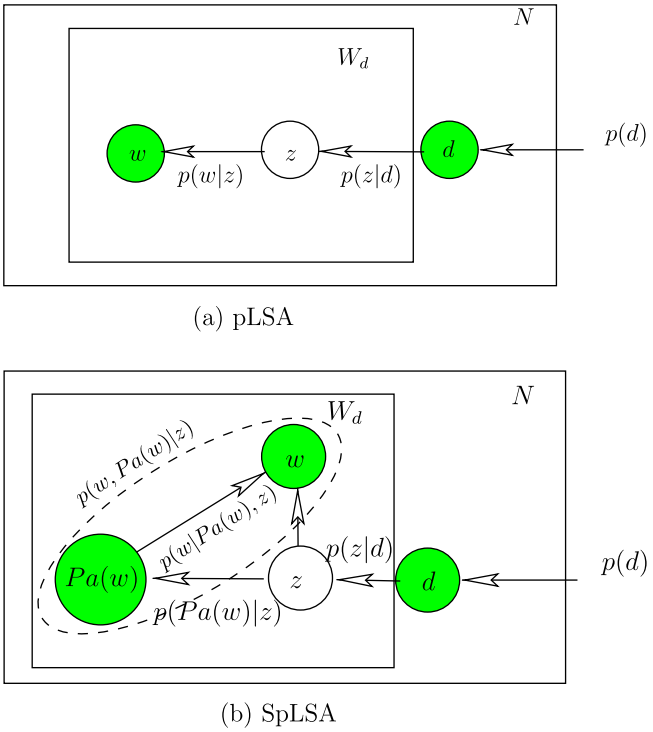

Fig. 1. Graph models of the pLSA (a) and SpLSA (b) respectively. Nodes are random variables. Filled nodes are observed and white ones are hidden. Arrows represent the dependencies between nodes. In (b), the joint probability $p(w,Pa(w)|z)$ can be written in terms of a product: $p(w|Pa(w),z)p(Pa(w)|z)$.

$$p(w_1, w_2, \ldots, w_T, d) = p(d)p(w_1, w_2, \ldots, w_T|d) = p(d)\prod_{i=1}^{T} p(w_i|Pa(w_i), d) = p(d)\prod_{i=1}^{T} \sum_{z\in Z} p(w_i|Pa(w_i), z)p(z|d), \quad (3)$$

where $Pa(w_i)$ is the parent set of word w_i, excluding z for clear explanation, because z is the parent of all of the words within the document.

Note that different definitions of $Pa(w_i)$ result in modeling different types of local dependencies. If $Pa(w_i) = \{w_j\}$, that means w_i has only one parent w_j and we term the resulting model as the *first-order structural pLSA*. If $Pa(w_i) = \{w_j, w_k\}$, the resulting model can be called the *second-order structural pLSA*, and so on.

3.1. Model learning

Since the model contains latent variables z, it is straightforward to learn the parameters of the model by maximizing the likelihood function with the Expectation–Maximization algorithm. The procedure is quite similar to the learning process of the traditional pLSA. Given a set of D documents, the likelihood function of the structural pLSA is defined as follows:

$$L = \prod_{m=1}^{D} p(w_1, w_2, \ldots, w_{T_m}, d_m) = \prod_{m=1}^{D} p(d_m) \prod_{i=1}^{T_m} p(w_i|Pa(w_i), d_m), \quad (4)$$

where T_m is the length of the sequence d_m (here we consider the representation of each clip as a word).

In the E step of the EM algorithm, the expectation of the posterior probability of the latent topic z_k is computed based on the Bayesian rule as follows:

Please cite this article in press as: J. Zhang, S. Gong, Action categorization by structural probabilistic latent semantic analysis, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.04.006

$$p(z_k|d_m, w_i, Pa(w_i)) = \frac{p(w_i, Pa(w_i)|z_k)p(z_k|d_m)}{\sum_{l=1}^{K} p(w_i, Pa(w_i)|z_l)p(z_l|d_m)}, \quad (5)$$

where K is the number of topics. The M step simply updates the following equations:

$$p(w_j, Pa(w_j)|z_k) = \frac{\sum_{m=1}^{D} n(d_m, w_j, Pa(w_j))p(z_k|d_m, w_j, Pa(w_j))}{\sum_{j=1}^{W}\sum_{m=1}^{D} n(d_m, w_j, Pa(w_j))p(z_k|d_m, w_j, Pa(w_j))}, \quad (6)$$

$$p(z_k|d_m) = \frac{\sum_{j=1}^{W} n(d_m, w_j, Pa(w_j))p(z_k|d_m, w_j, Pa(w_j))}{n(d_m)}. \quad (7)$$

It is worth to note that the term $p(w_i,Pa(w_i)|z)$ represents the joint probability of the co-occurrence of a word and its neighbors. The term $n(d_m,w_j,Pa(w_j))$ denotes the number of times the term w_j and its parents $Pa(w_j)$ co-occurred in the document d_m. Thus it has the ability of modeling temporal dependencies between words when applied to action recognition. We can consider the proposed model as an unsupervised version of the hidden Markov model without knowing the document (sequence) labels.

4. Structural pLSA for action recognition

4.1. SpLSA for actions

Actions are a time sequence, where temporal dependencies are one of the key characteristics. To capture such local temporal dependencies, a simple way in a computationally tractable manner is to assume Markov property, i.e., the word w_i at time t depends only on previous one w_{i-1} at time $t-1$, i.e., $Pa(w_i) = \{w_{i-1}\}$ (the words are now ordered in time, i.e., word w_{i-1} occurred earlier than w_i). By incorporating such dependencies, the first-order SpLSA can be further written as a compact version of Eq. (3):

$$p(w_1, w_2, \ldots, w_T, d) = p(d)\prod_{i=1}^{T}\sum_{z\in Z} p(w_i|w_{i-1}, z)p(z|d). \quad (8)$$

Accordingly, the likelihood function of Eq. (4) can be written:

$$L = \prod_{m=1}^{D} p(w_1, w_2, \ldots, w_T, d_m) = \prod_{m=1}^{D} p(d_m)\prod_{i=1}^{T} p(w_i|w_{i-1}, d_m). \quad (9)$$

It is worth to note that the mathematical form of the first-order SpLSA model (Eq. (8)) for an action sequence is quite similar to the hidden Markov model if we consider w as a hidden state. Both of them model the temporal dependencies between consecutive nodes. A remarkable difference is that first-order SpLSA has an additional latent variable representing topics z. Another difference is that, when learning HMMs for multiple actions, the action label of each sequence usually has to be known. We will compare their performance in the experimental section.

4.2. Recognition

Determining the model parameters involves fitting the SpLSA to the entire set of the training samples. During this procedure, we can learn the topic specific distributions $p(w,Pa(w)|z)$. Each training sample can be represented by a Z-vector $p(z|d_{train})$ where Z is the number of topics learnt. In our experiments, we set the number of topics as the number of action categories in a similar way to [25]. Note that the training process is totally unsupervised, i.e., it is not necessary to supply an action label to each sequence or any segmentation.

For an unseen sequence d_{test}, we model it as the conditional distribution of a mixture of topics, i.e., $p(z|d_{test})$. It is computed with EM fold in a similar manner to other pLSA based approaches described in [18,25,31,41]. In principle, the unseen sequence is projected onto the simplex spanned by the $p(w,Pa(w)|z)$ learnt during training, i.e., the mixing coefficients $p(z|d_{test})$ are sought such that the Kullback–Leibler divergence between the measured empirical distribution and $p(w,Pa(w)|d) = \sum_z p(w,Pa(w)|z)p(z|d_{test})$ is minimized. This is achieved by running EM in a similar manner to that used in training, but only the coefficients $p(z_k|d_{test})$ are updated in each M step with learnt $p(w,Pa(w)|z)$ kept fixed. The result is that each test sequence is represented by a Z vector. We then deploy a nearest neighbor classifier to classify the test samples against the training set.

4.3. Action features

In our experiments, we are interested in utilizing a feature representation for silhouette-based action recognition given images captured from fixed camera views with stable and/or known background information. The local shape context descriptor [1] has shown its promise for shape description. Inspired by this, our local feature description is in principle the same as the local shape context. On the other hand, due to the great success of the bag of visual words in generic object/texture categorization [42,10] (robustness to change of view points, clutters, intra-class variation), we are motivated to build a global descriptor in terms of a bag of local visual shapes. The feature extraction process is described as follows:

(1) Local shape context modeling. For each human silhouette, we randomly sample n points on the shape. For each point, we determine the local support region and extract the local shape context features according to the method described in [1]. The dimensionality of the local shape descriptor for each point on the shape is 60. The number of sample points for each shape is 100. The distances between those sample points are normalized by their mean value to achieve scale invariance.
(2) Codebook construction. We randomly select a number of action sequences from the training set and use the k-means algorithm to cluster the descriptors of all of the silhouettes from the training set within each action separately and then concatenate those cluster centers together to form a final codebook. The number of cluster centers per action is selected experimentally. In order to discriminate this with the 'words' used in the experiment, we term the codebook here as the *shapelet* codebook, i.e., each code is a representation of a certain type of shape parts. An example representation of the codebook is shown in Fig. 2. We have also tried the much more rigors method GMM for codebook construction, however, we find that in practical the k-means gives comparable or even better results than GMM. Possible reasons are the learning of GMM involves estimating a set of parameters in a very high dimensional space and usually leads to unstable solutions due to the relatively small number of the training samples. Due to the simplicity and the lower computational cost of the k-means against GMM, we adopt the k-means for codebook construction in all of our experiments.
(3) Codebook histogram. For each image from either training or test set, we construct a *shapelet* codebook histogram as a global descriptor, where each entry is the frequency of a certain *shapelet* prototype occurred in a silhouette image. Finally every histogram is normalized to have an unit $L1$ norm.

Fig. 2 shows each step of the feature extraction process.

Please cite this article in press as: J. Zhang, S. Gong, Action categorization by structural probabilistic latent semantic analysis, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.04.006

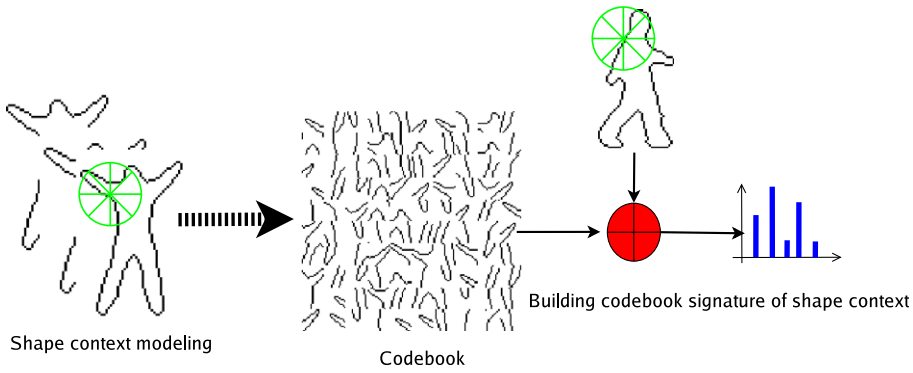

Fig. 2. Illustration of the action feature extraction process based on the local shape context descriptors.

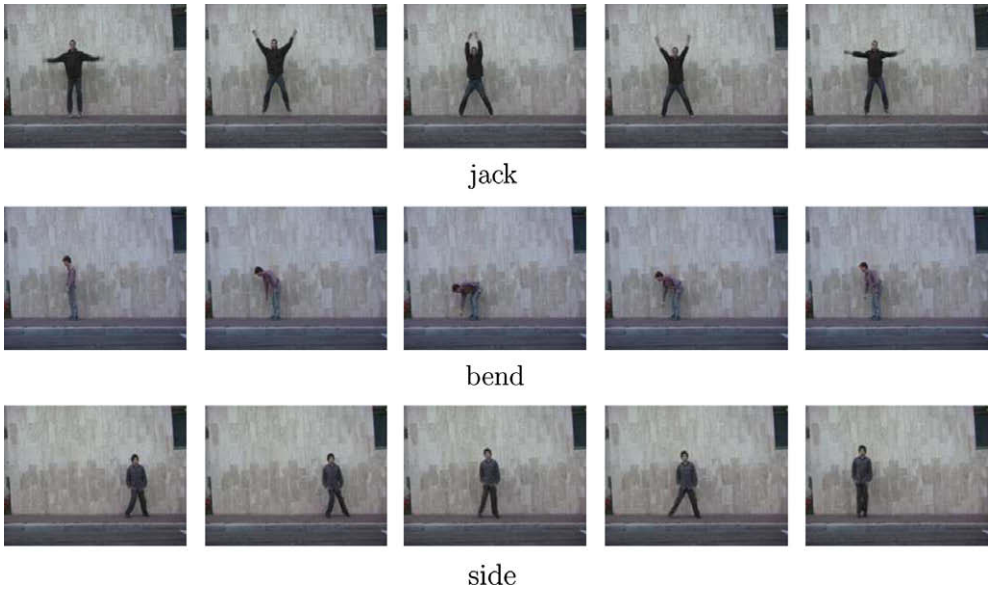

Fig. 3. Image samples of some action categories from the WEIZMANN dataset.

5. Experiments

5.1. Datasets

We evaluated our model on two different challenging datasets: (i) Blank et. al [2] and (ii) Wang et. al [38]. We will refer to these datasets as the WEIZMANN and MIT-CSAIL datasets respectively.

5.1.1. WEIZMANN dataset

This dataset is from [2]. It 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 such as *'running'*, *'walking'*, *'jumping-jack'*, *'jumping-forward-on-two-legs'*, *'jumping-in-place-on-two-legs'*, *'galloping-sideways'*, *'waving-two-hands'*, *'waving-one-hand'*, *'bending'*, *'skipping'*. We use all of the 10 classes in our experiment.[2] Fig. 3 shows some example action images from this dataset. Similar to [2], the silhouette of each frame is extracted based on the subtraction of the median background from that frame followed by a simple thresholding operation in color-space. The resulting silhouettes contain 'leaks' and 'intrusions' due to imperfect subtraction, shadows and color dissimilarities with the background. The local shape context feature is extracted from each silhouette image and 10 action sequences are randomly selected to construct the codebook with 20 cluster centers created per action sequence. Thus, the shapelet codebook histogram of 200 dimensionality is generated to describe each frame according to Section 4. In modeling the action sequences, we need to produce the words mentioned in pLSA or SpLSA (Here each action sequence corresponds to one document, and each frame corresponds to one word. Thus for an action sequence, the word instances are temporally ordered). To do this, we use the *k*-means to further cluster the shapelet histograms of all of the frames into a set of clusters(words). To test the robustness of our method, we have tried different number of words (see Table 3 for information). We use half of the sequences for training and the rest for testing. For the experiments, the data split is performed in a way that the testing dataset has no participants from the training set.

5.1.2. MIT-CSAIL dataset

This dataset is used in [38]. It includes six classes of arm gestures: *'Expand Horizontally'*, *'Expand Vertically'*, *'Point and Back'*, *'Double Back'*, *'Flip Back'*, and *'Shrink Vertically'*. These gestures are performed by thirteen users in front of a stereo camera and an average of 90 gestures per class were collected. For each image frame, a 3D cylindrical body model, consisting of a head, torso, arms and forearms was constructed using a stereo-tracking

[2] This is different from the settings of [2], where they only used 9 of them.

Please cite this article in press as: J. Zhang, S. Gong, Action categorization by structural probabilistic latent semantic analysis, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.04.006

algorithm [6]. From these body models, both the joint angles and the relative coordinates of the joints of the arms are used as observations for our experiments. Since the direct features are available for fair comparison with the results in [38], so feature extraction during prepossessing for these sequences is not necessary for this dataset. We keep the same training/test split as in [38].

5.2. Comparing different action models

In our experiments, we compared several existing model-based approaches for action recognition. They are described briefly as follows:

5.2.1. GMM

This is a generative model where each class is learnt as a Gaussian mixture model (GMM), i.e., $p(x;\Theta_c)$ for action class c, where x is a sample frame of an action sequence. For a frame x_t at time t, we assign its class label based on the maximization of a posterior probability, i.e., $c_t^* = \arg max_c p(\Theta_c|x_t)$. For a sequence within time T, the class label of the whole sequence is determined by a major voting strategy, i.e., $c^* = \arg max_c(p(c))$ with $p(c) = \frac{1}{T}\sum_{t=1:T}\delta(c_t^* = c)$. In our experiments, the number of mixture components is automatically determined by using the MDL criteria.

Note that GMM assumes the independence between local observations, thus it has no capability to model the temporal dependencies between consecutive frames.

5.2.2. 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 the nodes. The difference with GMM is that it is optimized based on the conditional probability given the sample labels. The final class label of the whole sequence is determined in a similar way to GMM.

5.2.3. SVM

It is one of the well established classifiers available today. Similar to the logistic regression, it is a discriminative method without the consideration of the dependencies between frames, but optimized by maximizing the separation margin between classes. We learn a multi-class SVM with RBF kernel, and the best parameters are learnt by cross validation. We then label each frame by the output of SVM. For a whole sequence, the sequence label is determined in a similar way to GMM.

5.2.4. HMM

This model is capable of modeling the temporal dependencies between hidden variables though it is a generative model. In our experiments, the number of hidden states and the transition matrix are automatically initialized by learning a GMM with the MDL criteria over the whole training set. We then learn a HMM for each class respectively, denoted by M_c. Thus for a given sequence $\mathbb{X}$, the class label is assigned via $c^* = \arg max_c p(M_c|\mathbb{X})$.

5.2.5. CRF

A conditional random field (CRF) is a discriminative model with the 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 the whole classes corresponding to each class label, and then infer the Viterbi path for each test sequence. The label of the whole sequence is computed as the most frequently happened frame label in the Viterbi path.

5.2.6. HCRF

A hidden conditional random field is an extension of the CRF by adding hidden state structures. It is optimized based on conditional probability of the action sequence label (instead of the node label in the case of CRF) given the observation of each node. Thus it naturally suits the task of action categorization. In our experiment, we learn HCRF using the EM algorithm.

5.3. Results

5.3.1. Comparison with pLSA

In order to test the performance of our method, we perform the experiment of action classification on the Weizmann and MIT-CSAIL dataset respectively. We investigate the performance of our method by varying the number of words. Our method is compared with the standard pLSA based on the same experimental settings. Table 3 tabulates the classification accuracy on the WEIZMANN dataset, whilst Table 4 gives the results for the MIT-CSAIL dataset. From this table, we can see that the performance of both pLSA and SpLSA is not a linear function of the number of words, while SpLSA is more robust to the change of number of words.

Overall, SpLSA outperforms pLSA and is significantly better than pLSA when the number of words is small, e.g., performance increases by 14.4% for 10 words and by 2.6% for 200 words on the MIT-CSAIL dataset; performance increases by 2.5% for 10 words and by 0.2% for 200 words on the WEIZMANN dataset. This might indicate that when the number of words is small (less words often means less informative, since a lot of data points with differences have been quantized into one word), the word dependencies become significant in the classification. This is somewhat similar to the observations in object categorization [10,42,21], i.e., pyramid models with spatial constrains tend to give better results than orderless bag of words model when the number of words is small, while their performances become comparable on many datasets with a large number of words.

In order to examine the performance on each individual category, we further give details of the confusion matrix of using SpLSA on the two datasets in Tables 1 and 2 respectively. From Table 1, we can see that 8 out of the 10 categories have the classification accuracy of 100%. The error is caused by the wave 1 and wave 2 categories. For the MIT-CSAIL dataset, 4 out of 6 categories have the accuracies higher than 93%, while the most difficult categories are FB and SV. FB is confused more often with EV and PB, while SV is mostly confused with FB. It is also interesting to note that the performance gain of SpLSA against pLSA on the MIT-CSAIL dataset is higher than the one on the WEIZMANN dataset. This indicates that temporal information (dynamics) plays a more important role for the MIT-CSAIL dataset than the WEIZMANN dataset, where the static action shape information of the actor is the key player for discrimination.

It is worth pointing out that TSI-pLSA [12] and pLSA-ISM [41] have also been proposed as extensions of pLSA. However, those models are developed based explicitly on the use of spatial information, in particular the coordinates of a local 3D patch, denoted as x_{rel} in [41]. On one hand, this enables a model to localize the action region. But on the other hand, it also limits such models to 3D local patch based action representations. If spatial information is not readily available, e.g., in the case of body shape utilized in our approach, these models would have been reduced to a standard pLSA, which we have compared with using the identical feature set discussed in this section.

5.3.2. Comparison with other existing models

We further compare the performance of our approach with other existing models as described in Section 5.2 available today. Tables 5 and 6 show the classification results on the WEIZMANN

Please cite this article in press as: J. Zhang, S. Gong, Action categorization by structural probabilistic latent semantic analysis, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.04.006

Table 1
confusion matrix of action classification results with SpLSA on the WEIZMANN dataset. The term 'jack' represents 'jumping-jack', 'pjump' for 'jumping-in-place-on-two-legs', 'side' for 'galloping-sideways', 'wave 1' for 'waving-one-hand' and 'wave 2' for 'waving-two-hands'.

Action	Bend	Jack	Jump	Pjump	Run	Side	Walk	Wave 1	Wave 2	Skip	
Bend	4										1.000
Jack		4									1.000
Jump			4								1.000
Pjump				4							1.000
Run					5						1.000
Side						4					1.000
Walk							5				1.000
Wave 1	2							2			0.500
Wave 2		1							3		0.750
Skip										5	1.000

Table 2
Confusion matrix of action classification results with SpLSA on the MIT-CSAIL dataset.

Action	EH	EV	PB	DB	FB	SV	Rate
EH	41	2		1			0.932
EV	2	56		1			0.949
PB			56	2	1		0.949
DB			1	65			0.985
FB	1	11	5		72	1	0.800
SV					5	40	0.889

Table 3
Classification accuracy of action categories using SpLSA with different number of words on the MIT-CSAIL dataset.

# Words	10	50	100	200
pLSA	0.729	0.851	0.851	0.774
SpLSA	0.873	0.882	0.909	0.800

Table 4
Classification accuracy of action categories using SpLSA with different number of words on the WEIZMANN dataset.

# Words	10	50	100	200
pLSA	0.877	0.923	0.921	0.898
SpLSA	0.902	0.930	0.930	0.900

Table 5
Classification accuracy of action categories using different models on the MIT-CSAIL dataset.

Models	GMM	LR	SVM	CRF	HMM	HCRF	SpLSA
Accuracy	0.719	0.702	0.747	0.868	0.868	0.915	0.909

Table 6
Classification accuracy of action categories using different models on the WEIZMANN dataset.

Models	GMM	LR	SVM	CRF	HMM	HCRF	SpLSA
Accuracy	0.861	0.899	0.930	0.930	0.907	0.931	0.930

and MIT-CSAIL dataset respectively. From the comparison of the results on the MIT-CSAIL dataset as shown in Table 5, we can see that the graph models with modeling temporal information give much better results than those graph models without modeling the temporal information, i.e., CRF, HMM, HCRF, and SpLSA give significant higher accuracy than GMM, LR, and SVM. Discriminative methods without modeling temporal information are not a clear advantage, i.e., SVM and LR perform comparably to GMM. However, *discriminative* learning of the latent structure using HCRF is an advantage against the *generative* learning approach using HMM. It is surprising to note that the *unsupervised* SpLSA performs better than the *supervised* models of HMM and CRF, and gives comparable results to the advanced HCRF approach. This clearly demonstrates the efficacy of the proposed approach.

For the WEIZMANN dataset, we can see that discriminative methods give better results than generative methods, i.e., SVM and LR produce higher accuracy than GMM; HCRF and CRF perform better than HMM. Note that HMM gives slightly better results than GMM, which shows that temporal information is not a very significant cue for this dataset. Using temporal information is not a clear advantage for discriminative learning methods, i.e., HCRF and CRF perform comparably to SVM and LR. The proposed method gives good results on this dataset, slightly better than HMM and comparable to HCRF.

6. Discussion and conclusions

In this work, we have presented a new model called *structural pLSA* to model the word dependency in a document. Accordingly, we have developed a novel action categorization approach using SpLSA with a codebook representation of actions by the local shape context features. Compared to other models, the learning process of the approach is *unsupervised*. Results on two challenging datasets show that the performance of SpLSA is superior to pLSA, especially when the number of words is small. The extensive comparison with other existing models indicates that the proposed approach achieves comparable even better results than the supervised learning approaches such as HMM, noticeably, comparable results to the state-of-art approach using HCRF with exactly the same feature set.

Though silhouettes extraction from still image segmentation is still a fundamentally challenging task in computer vision, this task can be simple and easier in many scenarios with fixed cameras installed in real surveillance applications. In these cases, the appearance of the background is known and a simple background extraction method usually results in satisfactory segmentation [15]. In more challenging conditions with the presence of illumination changes as well as some type of background motion clutter, the adaptive background modeling approach [34] can be employed to produce good extraction of human shapes.

In a non-stationary camera setting involving a moving camera or from a PTZ camera such as some of the action examples captured in the KTH dataset [30], auto-extract shape data reliably becomes challenging due to significant variations in 3D pose, the speed of motion and distance to the camera from the actions captured, resulting in significant shape variations. Some degree of manual correction/semi-supervised labeling is needed. This process poses a major undertaken in effort even with well designed background extraction models, e.g., adaptive GMM model [34], code book model, or kernel density based model. Thus an interest

Please cite this article in press as: J. Zhang, S. Gong, Action categorization by structural probabilistic latent semantic analysis, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.04.006

points based framework [30] is more suitable to this type of dataset. For future work, it would be interesting to exploit how to combine our approach with a 3D local patch based approach, and to evaluate the performance of such a combined model against data captured by non-stationary cameras.

Our model is temporal-shift invariant since the dependencies do not wary with time. Currently there is no special theoretical treatment about adaptation to different temporal resolutions. In practice, the temporal resolution adaptation can be achieved by training the model with samples of different temporal scales. Those samples can be created by re-sampling the original sequences under different temporal resolutions. It is worth noting that another popular trend for action recognition is of using spatial–temporal local interest points. The future work could involve developing hybrid feature representation to incorporate a global shape descriptor and features of local spatial–temporal volumes. Extending the current approach for action detection would be another valuable research direction.

Acknowledgment

The authors would like to thank Sybor Wang of MIT-CSAIL Lab for providing their gesture datasets and share their experimental settings in [38].

References

[1] 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.
[2] 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.
[3] 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.
[4] O. Boiman, M. Irani, Detecting irregularities in images and in video, International Journal of Computer Vision 74 (1) (2007) 17–31.
[5] A. Bosch, A. Zisserman, X. Muñoz, Image classification using ROIs and multiple kernel learning, International Journal of Computer Vision, (2008), submitted for publication.
[6] D. Demirdjian, T. Darrell, 3-D articulated pose tracking for untethered diectic reference, in: IEEE International Conference on Multimodal Interfaces, 2002, pp. 267–272.
[7] A. Demiriz, K.P. Bennett, Linear programming boosting via column generation, Journal of Machine Learning 46 (1–3) (2002) 225–254.
[8] P. Dollar, V. Rabaud, G. Cottrell, S. Belongie, Behavior recognition via sparse spatio-temporal features, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005, pp. 65–72.
[9] 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.
[10] M. Everingham, A. Zisserman, C.K.I. Williams, L. Van Gool, et al., The 2005 pascal visual object classes challenge, in: J. Quinonero-Candela, I. Dagan, B. Magnini, F. dÁlche-Buc (Eds.), Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment, Lecture Notes in Artificial Intelligence, vol. 3944, Springer-Verlag, 2006, pp. 117–176.
[11] L. Fei-fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories, Computer Vision and Image Understanding 106 (1) (2007) 59–70.
[12] R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from google's image search, in: International Conference on Computer Vision, 2005, pp. 1816–1823.
[13] V. Ferrari, L. Fevrier, F. Jurie, C. Schmid, Groups of adjacent contour segments for object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (1) (2008) 36–51.
[14] K. Fukunaga, Introduction to Statistical Pattern Recognition, second ed., Academic Press Professional, Inc., San Diego, CA, USA, 1990.
[15] L. Gorelick, M. Blank, E. Shechtman, M. Irani, R. Basri, Actions as space–time shapes, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (12) (2007) 2247–2253.
[16] G. Griffin, A. Holub, P. Perona, Caltech 256 Object Category Dataset, Technical Report ucb/csd-04-1366, California Institute of Technology, 2007.
[17] 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, pp. 1117–1120.
[18] T. Hofmann, Unsupervised learning by probabilistic latent semantic analysis, Machine Learning 42 (1–2) (2001) 177–196.
[19] I. Laptev, On space–time interest points, International Journal of Computer Vision 64 (2) (2005) 107–123.
[20] I. Laptev, B. Caputo, C. Schüldt, T. Lindeberg, Local velocity-adapted motion events for spatio-temporal recognition, Computer Vision and Image Understanding 108 (3) (2007) 207–229.
[21] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 2006, pp. 2169–2178.
[22] B. Leibe, E. Seemann, B. Schiele, Pedestrian detection in crowded scenes, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2005, pp. 878–885.
[23] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60 (2) (2004) 91–110.
[24] T.B. Moeslunda, A. Hilton, 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.
[25] J.C. Niebles, H. Wang, L. Fei-Fei, Unsupervised learning of human action categories using spatial–temporal words, in: British Machine Vision Conference, vol. 3, 2006, pp. 1249–1258.
[26] S. Nowozin, G. Bakir, K. Tsuda, Discriminative subsequence mining for action classification, in: IEEE International Conference on Computer Vision, vol. 10, 2007, pp. 1919–1923.
[27] A. Opelt, A. Pinz, A. Zisserman, Learning an alphabet of shape and appearance for multi-class object detection, International Journal of Computer Vision 80 (1) (2008) 16–44.
[28] J. Pei, J. Han, B. Mortazavi-asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu, Mining sequential patterns by pattern-growth: the prefixspan approach, IEEE Transactions on Knowledge and Data Engineering 16 (11) (2004) 1424–1440.
[29] A. Quattoni, S. Wang, L.-P. Morency, M. Collins, T. Darrell, Hidden conditional random fields, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (10) (2007) 1848–1852.
[30] C. Schuldt, I. Laptev, B. Caputo, Recognizing human actions: a local svm approach, in: International Conference on Pattern Recognition, Cambridge, UK, 2004, pp. 32–36.
[31] J. Sivic, B. Russell, A. Efros, A. Zisserman, W. Freeman, Discovering objects and their location in images, in: IEEE International Conference on Computer Vision, 2005, pp. 370–377.
[32] J. Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: International Conference on Computer Vision, vol. 2, 2003, pp. 1470–1477.
[33] C. Sminchisescu, A. Kanaujia, D. Metaxas, Conditional models for contextual human motion recognition, Computer Vision and Image Understanding 104 (2–3) (2006) 210–220.
[34] C. Stauffer, W.E.L. Grimson, Adaptive background mixture models for real time tracking, in: IEEE Conference on Computer Vision and Pattern Recognition, vol. 2, 1999, pp. 246–252.
[35] 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, 2007, pp. 134–141.
[36] 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) 1505–1518.
[37] S. Wang, S. Wang, R. Greiner, D. Schuurmans, L. Cheng, Exploiting syntactic, semantic and lexical regularities in language modeling via directed markov random fields, in: Proc. Int. Conf. on Machine Learning, 2005, pp. 948–955.
[38] 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.
[39] Y. Wang, H. Jiang, M.S. Drew, Z.-N. Li, G. Mori, Unsupervised discovery of action classes, in: IEEE Conference on Computer Vision and Pattern Recognition, 2006, pp. 1654–1661.
[40] D. Weinland, R. Ronfard, E. Boyer, Free viewpoint action recognition using motion history volumes, Computer Vision and Image Understanding 104 (2–3) (2006) 249–257.
[41] S.-F. Wong, T.-K. Kim, R. Cipolla, Learning motion categories using both semantic and structural information, in: IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–6.
[42] J. Zhang, M. Marszaek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: a comprehensive study, International Journal of Computer Vision 73 (2) (2007) 213–238.

Please cite this article in press as: J. Zhang, S. Gong, Action categorization by structural probabilistic latent semantic analysis, Comput. Vis. Image Understand. (2010), doi:10.1016/j.cviu.2010.04.006