Human Action Recognition Using Supervised pLSA

Tingwei Wang[1, 2] and Chuancai Liu[1]

[1]*School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China*
[2]*SM, University of Jinan, Jinan 250002, China*
tingweiwang@163.com, chuancailiu@mail.njust.edu.cn

Abstract

Probabilistic latent semantic analysis (pLSA) has been widely used by researchers for human action recognition from video sequences. However, one of the major disadvantages of pLSA and its other extensions is that category labels of training samples are not fully used in model learning procedure for classification task. In this paper, a supervised pLSA (spLSA) model is proposed for overcoming this drawback. By adding an observable category variable to generative process of classic pLSA, spLSA is endowed with more discriminative power. Thus, this model provides a unified framework for semantic analysis and object classification, where the topics formulation is guided by spLSA towards more discriminative and the mapping between the topics and the action categories are described in a fully probabilistic manner. Experimental results show that spLSA substantially outperforms pLSA and achieves comparable or better performances than latent dirichlet allocation based supervised models and other state-of-the-art methods.

Keywords: *human action recognition; supervised pLSA; probabilistic graphical models; generative models*

1. Introduction

Human action recognition from video sequences is an active research topic in computer vision community [1]. Generative models [2-5], which express the complex relationships between observed and target variables, become popular for action recognition. Among them, those modelling temporal patterns of actions, such as HMM [2], suffer from very complex modeling due to large intra-class variations of human actions. However, Probabilistic Latent Semantic Analysis (pLSA) [4], which was originally developed for topics discovery in a text corpus, has shown promising results in terms of accuracy, robustness and simplicity. In this paper, we propose a novel pLSA based model for human action recognition.

For action representation, pLSA adopts bag-of-words (BoW) paradigm which consists of feature extraction and dictionary construction. In the published literature, the work done to feature extraction of human body falls into two categories: holistic representation [6, 7] and local representation [8-12]. The former encodes the human body as a whole and the latter regards it as a set of 2D patches or 3D blocks. One key advantage of the local representation against the holistic representation is that it can preserve most of the information for action recognition while accurate localization, background subtraction and tracking are not required. In this paper, we apply local representation methods as action descriptors.

As the most successful BoW based generative and topic models, both pLSA and LDA (Latent Dirichlet Allocation) have been deeply studied for classification task. Sivic *et al.*, [13] applied this model to discover the categories and the locations of objects in a set of unlabelled images. Fergus *et al.*, [14] extended pLSA to include spatial information in a translation and

scale invariant manner, and utilized this modified pLSA model to learn an object category just from its name. Niebles *et al.*, [15] applied pLSA and LDA to learn the probability distributions of the spatial-temporal words and the intermediate topics corresponding to human action categories. Wong [16] incorporated implicit shape model (ISM) into pLSA for 3D video analysis and action recognition. Zhang and Gong [17] proposed structural pLSA to model explicitly dynamic adjacent dependencies between words by introducing new latent variables. Wang and Mori [18] trained a LDA based model in a "semi-supervised" way for human action recognition from video sequences, where the latent topics in their models directly corresponded to class labels and some of the latent variables became observed. In the multi-class sLDA of Wang *et al.*, [19], the class labels were treated as global descriptors of the images and an approximate inference and estimation algorithm based on variational methods were derived. Zhu *et al.*, [20] applied max-margin learning method instead of maximizing the likelihood of data to train supervised LDA for prediction task. Krithara *et al.*, [21] extended pLSA to semi-supervised framework and presented two semi-supervised variants of the pLSA respectively with fake labels and a mislabeling error model. Recently, in the work by Wang *et al.*, [22], according to whether the word-aspect probability is directly used as the pLSA model parameter for classification, two supervised pLSA algorithms were proposed, where the latent aspects for training become observable variables and the initial values of the word-aspect probabilities are no longer randomly assigned. However, we find that the above modified pLSA models suffer from a limitation of being unable to discover enough semantic and discriminative information for action recognition. For example, [21] manually fixed the probability of category-topic as 1 or 0 according to whether the topic component belongs to a certain category and [22] specified the number of topics as the number of action classes and took the topic variables as observed ones.

To eliminate the limitation, we propose a supervised pLSA (spLSA) for action recognition in this paper. Specifically, we modify the pLSA in a supervised manner by adding an observable variable, which takes on values from the set of ground truth category labels, to generative process of classic pLSA. Moreover, since pLSA performs similarly to LDA in practice [13], it is interesting to compare the performance between supervised pLSA with supervised LDA. We assume that there is a probabilistic mapping between the original latent topics and the action categories. By learning these category-topic probabilities, spLSA is endowed with more discriminative power and provides a unified framework for semantic analysis and object classification, which results in more accurate recognition than pLSA.

The rest of the paper is organized as follows. Section 2 discusses our supervised pLSA model. The model fitting and classification methods of spLSA are proposed in Section 3. Section 4 presents our experimental results on KTH dataset, Weizmann dataset and HMDB51 dataset, and also compares our performance with other methods. Finally we conclude the paper and discuss the future work in Section 5.

2. Supervised pLSA

2.1. Classic pLSA

As the first widely used aspect model, pLSA takes the generative process for cooccurrences of words and documents as a probabilistic mixture model, where a latent unobserved variable is associated with each observation. We summarize pLSA briefly below in the context of video analysis and more details can be found in [4]. Suppose that $\boldsymbol{D} = \{d_1, ..., d_N\}$ is a set of video sequences with words from a dictionary $\boldsymbol{W} = \{w_1, ..., w_M\}$, and $\boldsymbol{Z} = \{z_1, ..., z_K\}$ is a set of unobserved topics which is also called

latent topics. The joint probability of the cooccurrence of word w_i and video d_j is defined by the mixture over latent topics:

$$P(d_i, w_j) = \sum_k P(d_i \mid z_k)P(w_j \mid z_k)P(z_k) \quad (1)$$

where $P(z_k)$ is the probability of the occurrence of latent topic z_k, $P(w_j \mid z_k)$ is the conditional probability of word w_j given topic z_k, and $P(d_i \mid z_k)$ is the conditional probability of video sequence d_i given topic z_k. The corresponding probabilistic graphical model is shown in Figure 1(a).

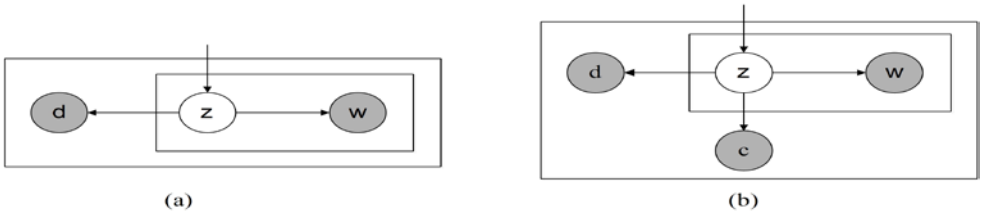

Figure 1. Graphical Models for (a) Classic pLSA and (b) our spLSA. Shaded Nodes are observed Variables and Unshaded ones are Unobserved Variables. The Plates Stand for Repetitions

2.2. Supervised pLSA

Though pLSA has achieved promising results in the application of automatic discovery of object categories, it ignores the important category information. To improve the discriminative power of pLSA, we add a category variable to its generative process. By following the previous definitions of $\boldsymbol{D}$, $\boldsymbol{W}$, $\boldsymbol{Z}$ in Section 2.1, let $\boldsymbol{C}$ be an action category variable that takes on values from the set of classes $\boldsymbol{C} = \{c_1, ..., c_L\}$. Let the category-topic distribution $P(c_l \mid z_k)$ be the conditional probability of category c_l given topic z_k. The generative process of our spLSA model for cooccurrences of video sequences, words, and categories is described as follows:

1. Choose a latent topic z_k according to probability $P(z_k)$,
2. Select a video d_i according to probability $P(d_i \mid z_k)$,
3. Generate a word w_j with probability $P(w_j \mid z_k)$,
4. Generate a category c_l with probability $P(c_l \mid z_k)$.

Essentially, this generative process implies video sequence d_i, word w_j and category c_l are statistically conditional independent given topic z_k. So the conditional joint distribution $P(d_i, w_j, c_l \mid z_k)$ can be denoted by $P(d_i \mid z_k)P(w_j \mid z_k)P(c_l \mid z_k)$. As a result, the joint probability of word w_j, video d_i, and category c_l can be written as follows:

$$P(d_i, w_j, c_l) = \sum_k P(d_i \mid z_k)P(w_j \mid z_k)P(c_l \mid z_k)P(z_k). \quad (2)$$

Figure 1(b) depicts the probabilistic graphical model of spLSA, where the above mentioned conditional independence between variables lead to the removal of lines from node d to c and from w to c in original full-linked probabilistic graph.

3. Model Fitting and Classification

3.1. Model Fitting

For model fitting of spLSA, we need learn the following parameters:

$$\Lambda = \{P(z_k), P(d_i \mid z_k), P(w_j \mid z_k), P(c_l \mid z_k) \mid k \in \{1,...,K\}, i \in \{1,...,N\}, j \in \{1,...,M\}, l \in \{1,...,L\}\}.$$

To estimate Λ, we apply maximum likelihood formulation to maximize the following log-likelihood of the complete data:

$$L = \sum_i \sum_j \sum_l n(d_i, w_j, c_l) \log P(d_i, w_j, c_l), \tag{3}$$

where $n(d_i, w_j, c_l)$ denotes the count of word w_j in video d_i with category label c_l. Due to the existing of latent topic z_k, there is not a analytical solution. Thus we use EM algorithm iteratively for model fitting, which is sketched out in Algorithm 1.

Given parameters Λ, E-step computes the posterior probability of each latent topic z_k at an iteration, by applying Bayes' formula:

$$P(z_k \mid d_i, w_j, c_l) = \frac{P(d_i \mid z_k)P(w_j \mid z_k)P(c_l \mid z_k)P(z_k)}{\sum_k P(d_i \mid z_k)P(w_j \mid z_k)P(c_l \mid z_k)P(z_k)}. \tag{4}$$

In the M-step, by using the method of Lagrange multipliers to maximize the expected complete data log-likelihood

$$E(L^c) = \sum_i \sum_j \sum_l n(d_i, w_j, c_l) \sum_k P(z_k \mid d_i, w_j, c_l) \log(P(d_i \mid z_k)P(w_j \mid z_k)P(c_l \mid z_k)), \tag{5}$$

we can get re-estimation equations:

$$P(z_k) \propto \sum_i \sum_j \sum_l n(d_i, w_j, c_l) P(z_k \mid d_i, w_j, c_l), \tag{6}$$

$$P(d_i \mid z_k) \propto \sum_j \sum_l n(d_i, w_j, c_l) P(z_k \mid d_i, w_j, c_l), \tag{7}$$

$$P(w_j \mid z_k) \propto \sum_i \sum_l n(d_i, w_j, c_l) P(z_k \mid d_i, w_j, c_l), \tag{8}$$

$$P(c_l \mid z_k) \propto \sum_i \sum_j n(d_i, w_j, c_l) P(z_k \mid d_i, w_j, c_l). \tag{9}$$

The E-step and M-step are repeated until the log-likelihood in Equation (3) converges, which is measured by a relative log-likelihood change between two successive EM runs. The time complexity of Equation (4, 5, 6, 7, 8, 9) are $O(K)$, $O(KNM)$, $O(NM)$, $O(ML)$, $O(N)$ and $O(NM)$, respectively. So the total time complexity of this algorithm is $O(KNM)$ which is equal to that of pLSA.

Algorithm 1.supervised pLSA

Input:

$n(d_i, w_j, c_l)$, where $i \in \{1,...,N\}$, $j \in \{1,...,M\}$, and $l \in \{1,...,L\}$,

The number of topics K

The max iteration number t.

1.Assign the parameters Λ randomly.

2.for $p = 1 \to t$

E-step: For $k \in \{1,...,K\}$, compute posterior probability of z_k by Equation (4).

M-step: Re-estimate parameters Λ by Equations (6,7,8 and 9).

If the log-likelihood in Equation (3) converges, break the 'for' loop.

end for

Output: The learnt parameters Λ

3.2. Classification

We cannot directly put a testing video into the above spLSA model due to the absence of category label. Alternatively, we adopt the fold-in scheme. Specifically, the testing sequence d_{test} is folded into a classic pLSA model, where the parameters $P(z)$ and $P(w|z)$ are fixed to the values learnt from spLSA. Then $P(z|d_{test})$ can be computed by $P(z|d_{test}) \propto P(d_{test}|z)P(z)$, where $P(d_{test}|z)$ is the output of the classic pLSA. Finally, the predictive label c^* of d_{test} is indicated as follows:

$$c^*(d_{test}) = \arg\max_l P(c_l|d_{test}) = \arg\max_l \sum_k P(c_l|z_k)P(z_k|d_{test}). \quad (10)$$

spLSA inherits the advantage of latent semantic discovery, which is represented by the $P(z|d_{test})$ distribution, from pLSA, and meanwhile bridges the gap between semantics and classification by the category-topic distributions $P(c|z)$. Thus, spLSA provides a unified framework for semantic analysis and object classification, where the topics formulation is guided by spLSA towards more discriminative and the mapping between the topics and the action categories are described by the category-topic distributions in a fully probabilistic manner. It is worth to be noted that our graphical model in Figure 1(b) is similar to the model proposed in [21], but there are fundamental differences between them. In theory, $P(c|z)$ are regarded as prior probabilities in [21], while they are model parameters in our spLSA. Consequently, [21] is time consuming for manual assignment of the values of $P(c|z)$, which results in an intractable problem when either the number of the latent topics increases or a large scale dataset is used. In contrast, these values can be automatically learnt by our spLSA in a probabilistic manner, which is more efficient.

4. Experiments and Results

We tested our model on three public datasets: KTH human motion dataset [23], Weizmann human action dataset [24], and HMDB51 dataset [25]. Example frames are shown in Figure 2.

4.1. Datasets

Figure 2. Example Frames from KTH (Top), Weizmann (center) and HMDB51(bottom)

KTH dataset is performed by 25 subjects. Each subject does six types of actions (walking, jogging, running, boxing, hand waving and hand clapping) under four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. Now, with one missed and one broken removed, there are 598 action video clips available. We adopted the original training set and testing set split method, *i.e.*, 9 subjects (2, 3, 5, 6, 7, 8, 9, 10, and 22) being training set and the rest 16 subjects being testing set.

Weizmann dataset contains ten types of natural human actions such as 'run', 'walk', 'jack', 'jump', 'pjump', 'side', 'wave1', 'wave2', 'bend' and 'skip'. Each action is performed by nine different persons. There are 93 video clips available. To increase the amount of samples of Weizmann, we flipped all the video sequences along the time axis, thus got a total of 180 video sequences. Leave-One-Out Cross-Validation (LOOCV) testing paradigm was used and average accuracy over all classes was reported.

To evaluate the scalability of our model, more realistic and challenging dataset, *i.e.*, HMDB51, was used to conduct experiments. HMDB51 dataset has been collected from a variety of sources ranging from digitized movies to YouTube. There are 51 distinct categories, each containing at least 101 clips for a total of 6,766 video clips. It is to-date one of the largest and most realistic available dataset. For evaluation purposes, three distinct training and testing splits were generated from the database. For each split, 70 clips of each category were selected as training set and 30 clips were selected as testing set. Our model was only evaluated on stabilized videos of HMDB51 and average accuracy over the three splits was reported.

4.2. Experimental Setup

To verify the robustness of our spLSA on different feature descriptors, we used three state-of-the-art spatial-temporal interest points (STIP) detectors and descriptors for the three datasets, respectively. Specifically, the code provided by Wang [12] for action representation based on dense trajectories and motion boundary histograms (MBH) descriptors was used for KTH, the code provided by Dollár [10] for Cuboid detection and descriptor was used for Weizmann, and HOG/HOF descriptors [25] provided by the dataset were applied for HMDB51. We adopted the BOW scheme to represent the action video, where the word numbers were fixed to 4000 and a subset of 100,000 features sampled from the training videos were clustered by K-means algorithm. As for every video clip, once each STIP descriptor is assigned to its nearest word, it is straightforward that they are represented as a histogram of these words occurrences.

The numbers of latent topics are regarded as model parameters and the optimal ones were got by exhaustively searching in a range (Given the category number L, ranging from L to $16*L$ with an increment of L). To eliminate the influence by initial values of EM, all experiments were conducted 5 times and the average accuracies were reported.

4.3. Comparison with other Topic Models

We compare our spLSA with pLSA+SVM, sLDA[1] and MedLDA[2] in Table 1 on the three datasets. pLSA+SVM uses the topic representation learn by pLSA as the feature vector of SVM. For SVM, we used the linear kernel and 5-fold cross validation on training set to select the optimal penalty parameter C from $\{k^2: k= -1,...,8\}$. Overall, our spLSA consistently outperforms the baseline pLSA by 1.5% to 10.7%.

Table 1. Recognition Accuracies (%) of our spLSA and other Topic Models

Method	KTH	Weizmann	HMDB51
spLSA	**95.8**	**99.3**	**23.6**
pLSA+SVM	93.9	97.8	12.9
sLDA	92.1	97.9	18.7
MedLDA	93.4	98.9	19.7

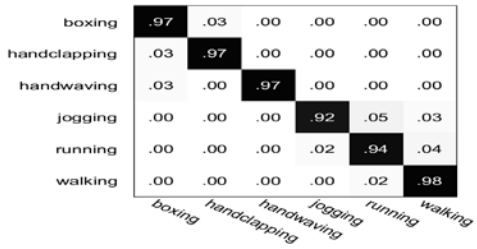

Figure 3. Confusion Matrix for KTH

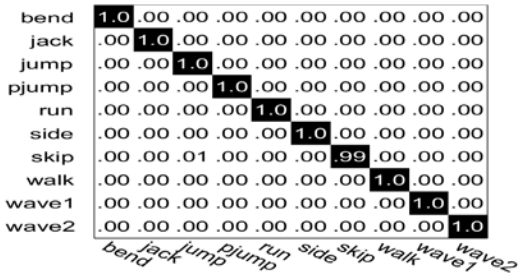

Figure 4. Confusion Matrix for Weizmann

spLSA outperforms pLSA by a small margin on Weizmann dataset, *i.e.*, 1.5%. This is because that this dataset is so small and simple that pLSA can distinguish the actions. The superiority of spLSA over pLSA is especially large on the realistic and challenging HMDB51, where spLSA is 10.7% better than pLSA. This confirms the advantage of capturing both semantic and discriminative information when training spLSA model. Though both sLDA and MedLDA adopt prior distribution of topics and word-topic, and complex variational inference, spLSA performs better than or comparable to these LDA-based supervised models. It is similar to the discovery that unsupervised pLSA performs similarly to unsupervised LDA in [13]. It is probably because that the prior dirichlet and multinomial distributions of LDA based methods fail to model exactly the action variations.

The confusion matrix for KTH of spLSA is depicted in Figure 3, which shows that the biggest classification errors are due to the similarity between 'jogging' and 'running'. Figure 4 presents the confusion matrix for Weizmann of spLSA, where the only confusion is made by the wrong assignment of 'skip' to 'jump'.

[1] Available at http://www.cs.cmu.edu/ chongw/slda/
[2] Available at http://www.ml-thu.net/ jun/medlda.shtml

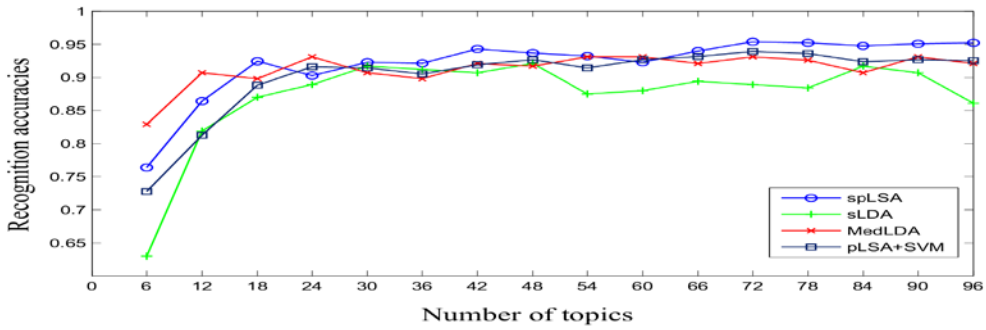

Figure 5. Recognition Accuracies vs. Topic Numbers on KTH

Figure 5 demonstrates the performance difference of all methods with respect to the number of topics on KTH dataset. It is observed that spLSA perform rather stably and achieves better accuracies than other topic models over almost all topics numbers except for several ones, *e.g.*, 24.

Table 2. Time Cost (second) of our spLSA and other Topic Models

method	KTH	Weizmann	HMDB51
spLSA	**9.8**	**4.5**	**280.1**
pLSA+SVM	10.4	4.8	290.5
sLDA	3355.2	262.3	8.46e+4
MedLDA	760.9	50.5	1.52e+4

It should be noted that our spLSA is efficient in running time. Table 2 lists the model training time cost (Intel CORE i5 Processor, 6G memory, and all methods are implemented in C/C++) of our spLSA and other topic models, where topic numbers are fixed to 5 times the action category numbers. In this experiment, the penalty parameter C of MedLDA was set to 32 for simplicity (the actual time cost may increase when larger C is used in practice). The fewest time cost of spLSA among these topic models means that our model is more suitable for real application. Although the computation of $P(c \mid z)$ are not required in the model fitting of pLSA+SVM, the whole training phrase of baseline pLSA+SVM is more time consuming than spLSA due to the cross validation for selecting optimal SVM parameter C. Due to the variational inference, the time cost of LDA-based methods are so high (*e.g.*, sLDA costs 8.46e+4 seconds on HMDB51) that become not very practical in large scale computer vision field.

Table 3. Recognition Accuracies (%) of spLSA and State-of-the-Art Methods

KTH		Weizmann		HMDB51	
spLSA	**95.8**	spLSA	**99.3**	spLSA	**23.6**
Wang et al.[12]	95.0	Wanget al.[22]	**100**	Kuehne et al.[25]	21.9
Wong et al.[16]	83.9	Zhang and Gong[17]	93	C2 [25]	23.1
Niebles et al.[15]	83.3	Niebles et al.[15]	90		

4.4. Comparison with state-of-the-art Methods

We further compare the performance of our model with other state-of-the-art methods which are topic model based methods or use the identical action descriptor to us. From the comparison in Table 3, we can see that spLSA achieves comparable or better recognition accuracies. Some of the comparisons are not precise due to the variations in the experimental setups. For example, [17] needs human detection processing and [22] requires human tracking and detection. However, our method does not need these preprocessing operations, which is possibly more practical for real applications.

On KTH and Weizmann, spLSA outperforms all pLSA based methods, *i.e.*, [15], [16] and [17], by a large margin, although [16] and [17] utilize spatial structure information. Because of the simplicity of Weizmann, [22] achieve 100% accuracies with the application of background segmentation and human tracking and stabilizing. The sPLA yields 99.3% which is comparable to them.

The most impressive comparison is the one on HMDB51 between [25] and our spLSA with the identical HOG/HOF feature set but the different classifier, *i.e.*, the powerful SVM for [25], where the latter outperforms the former by 1.7%. It shows that the discriminative method SVM has not a clear advantage against the spLSA when used on the large scale dataset captured from realistic environment.

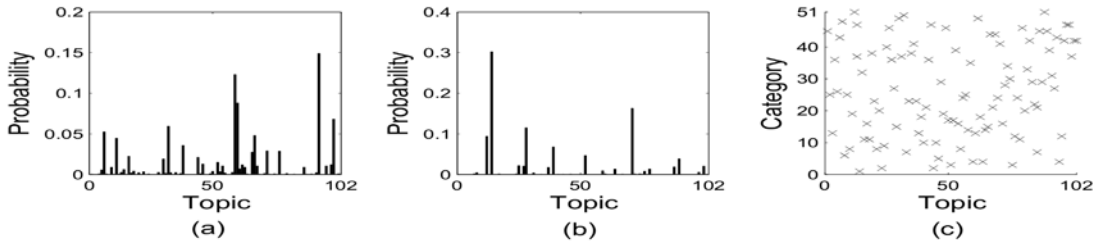

Figure 6. Visualization of Representative Model Parameters of spLSA and pLSA on HMDB51 (a) $P(z \mid d_{test})$ of pLSA, (b) $P(z \mid d_{test})$ of spLSA, (c) $P(c \mid z)$ of spLSA

4.5. Discussion

Visualizing the representative model parameters learnt by spLSA and pLSA in Figure 6, we find that there are three key components that make us to do better than pLSA. First, the video specific probability over topics $P(z \mid d_i)$ in spLSA (see Figure 6(b)) is more sparse than that of pLSA (see Figure 6(a)) by suppressing irrelevant latent topics. The refined sparse topics are guided by our spLSA towards more discriminative, resulting in a more valid capture of semantic and meaningful information. The second component is that our spLSA is flexible and allow the category-topic distributions to be described in a fully probabilistic fashion. Therefore, the mapping between topics and categories are 'soft', meaning that the values vary between 0 and 1. In fact, most of these category-topic probabilities learnt by spLSA are zero and a representative non-zero mapping are marked in Figure 6(c). This figure indicates that only certain topics, which can be automatically discovered by spLSA, contribute to a certain category of action. The last and most important, the many-to-many correspondences between topics and categories are the key factors that capture intra-class and inter-class variations and effectively bridge the gap between semantics and discrimination. For example, the topics contained in different 'sword' video sequences may vary according to

the changes of illumination, view and appearance, and in contrast, 'walk' and 'run' may consist of the same topic 'knee lift', which can be regarded as a minor pattern among actions.

5. Conclusion

In this work, we have presented a supervised pLSA model. By adding a category variable and describing the category-topic distributions in a fully probabilistic fashion, the novel spLSA is more discriminative than classic pLSA and its other extensions, and meanwhile still keeps the advantage of semantic analysis of pLSA. Although be proposed for human action recognition, spLSA can be easily used for other recognition tasks, such as document classification, object detection and scene understanding. We have verified the approach on three publicly available datasets. The experimental results show that spLSA substantially outperforms the baseline pLSA by a large margin and achieves better results than, or comparable ones to, several state-of-the-art methods, especially other topic models.

Our proposed model is still very simple and there are some future work to be done. To achieve better results, we will explore the scale-invariant spatial and temporal relationship between features under the spLSA framework. It is worth noting that another popular trend is combining max-margin learning scheme with topic models. Incorporating max-margin learning method into our spLSA would be of value in theory and practice.

Acknowledgements

The work was supported in part by National Natural Science Foundation of China (Nos. 60632050, 9082004 and 61202318) and National 863 Project of China (Nos. 2006AA04Z238, 2006AA01Z119).

References

[1] R. Poppe, "A survey on vision-based human action recognition", Image and Vis. Computing, vol. 28, no. 6, **(2010)**, pp. 976-990.
[2] J. Yamato, J. Ohya and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model", Proceedings of IEEE CS Conf. on Comput. Vis. Pattern Recognit., Champaign, IL, **(1992)**, pp. 379-385.
[3] Y. Luo, T. der Wu and J. Neng Hwang, "Object-based analysis and interpretation of human motion in sports video sequences by dynamic bayesian networks", Comput. Vis. Image Underst., vol. 92, **(2003)**, pp. 196-216.
[4] T. Hofmann, "Unsupervised Learning by Probabilistic Latent Semantic Analysis", Mach. Learning, vol. 42, no. 1-2, **(2001)**, pp. 177-196.
[5] D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation", J. of Mach. Learning Research, vol. 3, **(2003)**, pp. 993-1022.
[6] T. Kyun Kim and R. Cipolla, "Canonical Correlation Analysis of Video Volume Tensors for Action Categorization and Detection", IEEE Trans. on Pattern Anal. Mach. Intell., vol. 31, no. 8, **(2009)**, pp. 1415-1428.
[7] V. Maik, D. T. Paik, J. Lim and J. Paik, "Hierarchical pose classification based on human physiology for behaviour analysis", IET Comput. Vis., vol. 4, no. 1, **(2010)**, pp. 12-24.
[8] I. Laptev and T. Lindeberg, "Space-time interest points", Proceedings of IEEE Int. Conf. on Comput. Vis., Nice, France, **(2003)**, pp. 432-439.
[9] A. Oikonomopoulos, I. Patras and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions", IEEE Trans.on Systems, Man and Cybernetics, vol. 36, no. 3, **(2006)**, pp. 710-719.
[10] P. Dollár, V. Rabaud, G. Cottrell and S. Belongie, "Behavior recognition via sparse spatio-temporal features", Proceedings of Joint IEEE Int. Works on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, **(2005)**, pp. 65-72.
[11] I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, "Learning realistic human actions from movies", Proceedings of IEEE CS Conf. on Comput. Vis. Pattern Recognit., Anchorage, Alaska, USA, **(2008)**, pp. 1-8.

[12] H. Wang, A. Kl¨aser, C. Schmid and L. Cheng-Lin, "Action Recognition by Dense Trajectories", Proceedings of IEEE Conference on Computer Vision & Pattern Recognition, Colorado Springs, United States, **(2011)** June, pp. 3169-3176.
[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman and W. T. Freeman, "Discovering Objects and their Localization in Images", Proceedings of IEEE Int. Conf. on Comput. Vis., Beijing, China, **(2005)**, pp. 370-377.
[14] R. Fergus, F. fei Li, P. Perona and A. Zisserman, "Learning Object Categories from Google's Image Search", Proceedings of IEEE Int. Conf. on Comput. Vis., Beijing, China, **(2005)**, pp. 1816-1823.
[15] J. C. Niebles, H. Wang and F. Fei Li, "Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words", Int. J. Comput. Vis., vol. 79, no. 3, **(2008)**, pp. 299-318.
[16] S. fai Wong, T. kyun Kim and R. Cipolla, "Learning Motion Categories using both Semantic and Structural Information", Proceedings of IEEE CS Conf. Comput. Vis. Pattern Recognit., Minneapolis, Minnesota, USA, **(2007)**.
[17] J. Zhang and S. Gong, "Action categorization by structural probabilistic latent semantic analysis", Comput. Vis. Image Underst., vol. 114, no. 8, **(2010)**, pp. 857-864.
[18] Y. Wang and G. Mori, "Human action recognition by semi-latent topic models", IEEE Transactions on Pattern Analysis and Machine Intelligence Special Issue on Probabilistic Graphical Models in Computer Vision, vol. 31, no. 10, **(2009)**, pp. 1762-1774.
[19] C. Wang, D. M. Blei, and F. Fei Li, "Simultaneous image classification and annotation", Proceedings of Computer Vision and Pattern Recognition, **(2009)**, pp. 1903-1910.
[20] J. Zhu, A. Ahmed and E. P. Xing, "MedLDA: maximum margin supervised topic models for regression and classification", Proceedings of International Conference on Machine Learning, **(2009)**, pp. 158-1264.
[21] A. Krithara, M.-R. Amini, C. Goutte and J.-M. Renders, "Learning aspect models with partially labeled data", Pattern Recognit. Letter, vol. 32, no. 2, **(2011)**, pp. 297-304.
[22] J. Wang, P. Liu, M. F. She, A. Kouzani and S. Nahavandi, "Supervised learning probabilistic latent semantic analysis for human motion analysis", Neurocomputing, vol. 100, no. 1, **(2013)**, pp. 134-143.
[23] C. Sch¨uldt, I. Laptev and B. Caputo, "Recognizing Human Actions: A Local SVM Approach", Proceedings of Int. Conf. on Pattern Recognit., Cambridge, England, UK, **(2004)**, pp. 32-36.
[24] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri, "Actions as space-time shapes", Proceedings of IEEE Int. Conf. on Comput. Vis., Beiging, China, **(2005)**, pp. 1395-1402.
[25] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, "HMDB: a large video database for human motion recognition", Proceedings of IEEE Int. Conf. on Comput. Vis., Barcelona, Spain, **(2011)**, pp. 2556-2563.

Authors

Tingwei Wang received his B.S. degree in Computer Science and Technology from China University of Mining and Technology, China, in 2002 and M.S. degree in Computer Science and Technology from Qufu Normal University, in 2005, respectively. He is currently pursuing his Ph.D. degree in Computer Science from Nanjing University of Science and Technology. His research interests include action recognition and pose estimation. E-mail: tingweiwang@163.com.

Chuancai Liu is a full professor in the school of computer science and engineering of Nanjing University of Science and Technology, China. He obtained his Ph.D. degree from the China Ship Research and Development Academy in 1997. His research interests include AI, pattern recognition and computer vision. He has published about 50 papers in international/national journals. E-mail: chuancailiu@mail.njust.edu.cn.