Proceedings of the 2009 IEEE International Conference on Systems, Man, and Cybernetics
San Antonio, TX, USA - October 2009

Unsupervised Human Motion Analysis Using Automatic Label Trees

Kui Jia
Laboratory for Culture Integration Engineering
SIAT
Shenzhen, China
kuijia at gmail.com

Xie Wuyuan
Laboratory for Culture Integration Engineering
SIAT
Shenzhen, China
wy.xie at sub.siat.ac.cn

***Abstract*—Giving different human motions, there may be similar motion features in some body components, e.g., the motion features of arms when performing jogging and running, which are thus less discriminative for motion classification. In this paper, we consider counting less on these body components that have less discriminative information amongst different human motions. To this end, we present a new topic model, probabilistic latent semantic analysis based on multiple bags (M-PLSA), in which not all body components are considered equally important, i.e., motion features of less discriminative components are made less use of so that their contributions for classification are reduced. We use sparse spatio-temporal features extracted from videos to create visual words which are later assigned to different body components that they are detected from, so that co-occurrence matrices of different components can be calculated based on their corresponding vocabularies. Such label task can be automatically fulfilled by using the query visual words, i.e., words whose component labels are unknown, to traverse an automatic label tree (ALT) that grows from the training words with component labels. We show the performance of our approach on KTH dataset [1].**

***Index Terms*—topic model, component vocabulary, ALT**

I. Introduction

It is of great practical interests to analyze human motions in an unsupervised way, especially when access to large amount of training samples is out of the question. In motion understanding, one key problem is to extract discriminative motion features from videos so that the motions can be represented in an abstract and compact way. In [3] [4] [5], silhouette and stick figure were used as features to estimate human poses. Dollár et al. [6] used spatio-temporal features as visual words to represent videos, their experiments give good classification results for mouse motions and face expressions. Niebles et al. [7] went a step further by applying a constellation model to classify those videos represented by spatio-temporal interest points, their model is of a sort of topic model which involves relative position information between feature points. Those above methods treat all interest features extracted from a video as a whole set of video descriptors without body component labels, i.e. label of the component that an interest feature extracted from. However, giving two different motions, components with less discriminative information may exist, for example Fig 1, jacking and hand waving may have the same interest features of hand, so "hand" is a less discriminative component between them. Consequently, those models that count all detected features equally during motion classification may get less accurate results.

Our paper addresses the discriminative capabilities of different components to classify motions in a topic model. We present a novel topic model, probabilistic latent semantic analysis based on multiple bags (M-PLSA), in which each body component has its own independent vocabulary. Our model is inspired by Fei-Fei et al. [7] which is a hierarchical model based on multiple relative bag layers. Giving an input video, co-occurrence matrices associated with these vocabularies are computed, and each element of the matrices indicates the occurrence frequency of a particular interest feature detected from the corresponding body component. By using multiple co-occurrence matrices as input of M-PLSA, less discriminative components can be counted less in (5) (relative to the others), and their contribution for classification is reduced in turn. Experiment of our approach on KTH dataset [1] is given in Section 3.

On the other hand, to create visual words from extracted interest feature points, K-Means classification are used extensively [8] [9] [10] [11]. Some recent studies based object recognition in image [12] [13] [14] have resorted to random approaches for coding visual words. For example, Moonsmann et al. [12] used random forests as visual code books to code spatial feature points detected from images. Also, [15] [16] show better performance of random trees (RT) and random forests (RF) over K-Means in the aspect of speed as well as in the aspect of quality. Our approach pushes this line further. We use some *component-labeled* spatio-temporal interested points (STIP) [17], i.e., those STIP with the labels of the components that they are detected from, to grow a random tree, which is called automatic label tree (ALT) here. Different from RT and RF [13] [14] whose leaves are class labels and used for voting the class of a giving object, in ALT all those leaves with the same component label are treated as visual words from the vocabulary associated with this component. When traversing ALT by using a set of STIP extracted from a test video, co-occurrence matrices of each component can be computed, furthermore, the whole body can be automatically segmented into different components according to the resulting labels of STIP.

In summary, we show a multiple topic model M-PLSA that

978-1-4244-2794-9/09/$25.00 ©2009 IEEE

(a) hand waving

(b) jacking

Fig. 1. Example frames from data set [2].

has multiple independent "bags of words" with each word having been labeled automatically through a novel random tree ALT. Our contributions are three folds: First, M-PLSA is such a discriminative topic model that counts less on the body components which share similar motion features among different motion classes; Secondly, by traversing ALT, one can directly and quickly get co-occurrence matrices of visual words from the raw detected STIP whose descriptor dimension is not reduced; Finally, ALT can also be applied to segment body components, and is useful for motion tracking.

The rest of our paper is organized as follows. In Section 2, our theoretical framework is described in more detail, including brief introduction of topic model, random trees and our categorization procedure. In Section 3, experimented results on data set [1] are shown as well as performance comparison with Fei-Fei et al [7] method. Finally, Section 4 concludes this paper.

II. Theoretical Framework

Our model is a topic model based on multiple vocabularies. We segment the whole body into P components, e.g., hand, arm, foot and leg etc. Each component $p, (p \in 1 \dots P)$ has its own vocabulary $\mathbf{V}_p$ with size W_p, while each word $\omega_j^p, (j = 1 \dots W_p)$ from $\mathbf{V}_p$ is the descriptor vector for a spatio-temporal interest point detected at the pth component. We follow Dollár et al.'s [6] way to detect spatio-temporal cuboids and represent them by the gradient descriptor. Note that detected features using [17] [6] have no component information, we cannot tell which feature point is extracted from a particular body component. Such problem can be handled by using a descriptor vector to traverse ALT (cf. Section 2). In other words, component vocabularies are supposed to be independent with each other without losing geometric constraints as a whole. Actually, those component labels contain geometric information as they localize features. On the other hand, instead of giving the iteration accuracy to control the running time of EM algorithm (cf. Section 2) used in our model, we use topic unique constraints (TUC), i.e. each body component in a frame reflects the same motion class, as the stop criterion of the EM algorithm. We believe by doing so, classification error can be reduced.

A. The Multiple Bags Model

Given a video d, we find P sets of observed features $\{\omega_j^p | j \in (1 \dots W_p)\}$,$p \in (1 \dots P)$. Similarly, given a video set $\mathbf{D} = \{1 \dots N\}$ that contain K unknown motion classes $k \in (1 \dots K)$, we find P sets for each video $i, i \in (1 \dots N)$, then there are $N \times P$ sets of features in total. Based on these feature sets and giving P vocabularies $\mathbf{V}_p, p \in (1 \dots P)$, we can compute P co-occurrence matrices $\mathbf{M}_p, p \in (1 \dots P)$ for video set $\mathbf{D}$. The pth matrix $\mathbf{M}_p$ has the size of $N \times W_p$, where W_p is the size of the pth vocabulary. And an element $\mathbf{M}_p(i, \omega_j^p)$ of $\mathbf{M}_p$ has the value of $n_p(i, \omega_j^p)$ which is defined as the times of occurrence of a word ω_j^p in the ith video giving a particular component p. For a particular video set $\mathbf{D}$, each matrix $\mathbf{M}_p$ has the same size in row (i.e. N, size of video set $\mathbf{D}$), but has different size in column according to the size W_p of each vocabulary. Such a setting is convenient to make comparison of discriminability between different components. Giving two videos i and $i^{'}$ recording two different motions respectively, if $\mathbf{M}_p(i, j) \approx \mathbf{M}_p(i^{'}, j), \forall j \in (1 \dots W_p)$, then the pth component is less discriminative between video i and video $i^{'}$.

Probabilistic Latent Semantic Analysis (PLSA) was initially proposed to deal with polysemy in text understanding [18] [19]. Some recent works [7] [8] have used it for image object localization or categorization. Here, we follow the notation used in [20] plus the introduced variable p, which stands for the label of the pth body component. Definitions are as below: $p(\omega_j^p|k)$ is the probability of the visual words ω_j^p occurring in a particular motion class k and detected from the pth component, while $p(k|i)$ is the probability of topic k occurring in video i. Then the joint probability $P(i, \omega_j^p)$ and the conditional probability $p(\omega_j^p|i)$ are assumed to have the following expressions according to [20]

$$P(d_i, \omega_j^p) = p(i)p(\omega_j^p|i) \tag{1}$$

$$p(\omega_j^p|i) = \sum_{k=1}^{K} P(k|i)p(\omega_j^p|k) \tag{2}$$

Our aim is to estimate $p(\omega_j^p|i)$ from a set of observation pair in the ith video $\mathbf{O}_i = \{(i, \omega_j^p) | \forall j \in (1 \dots W_p); \forall p \in (1 \dots P)\}$.

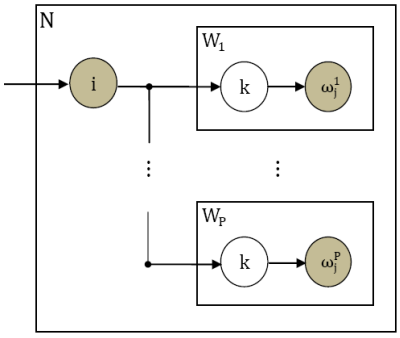


Fig. 2. The graphical model of M-PLSA. The outer plate indicates the graph is replicated for each video with N total ones, and the inner plate indicates the replication of the graph of each visual word for each component. The same common topic k is hidden among different body components.

According to [20], optimal values of $p(\omega_j^p|i)$ can be achieved by maximizing the objective function

$$L_i = \sum_{p=1}^{P}\sum_{j=1}^{W_p} n_p(i,\omega_j^p)\log P(i,\omega_j^p) \tag{3}$$

Actually, (3) takes the likelihood of each component into account, or we can say motion from each body component in video i is weighted with its frequency of feature occurrence $n_p(i,\omega_j^p)$ in (3). Likewise, for a video set $\mathbf{D} = \{1\dots N\}$, one can compute the maximum likelihood function L which sums over all $L_i, i \in (1\dots N)$. It yields the final form of object function

$$L = \sum_{i=1}^{N} L_i = \sum_{i=1}^{N}\Big(\sum_{p=1}^{P}\sum_{j=1}^{W_p} n_p(i,\omega_j^p)\log P(i,\omega_j^p)\Big) \tag{4}$$

Using the right item of equation (2) to replace $p(\omega_j^p|i)$ in equation (1), the variable of motion class k is introduced into L, and then by changing summing order in (4), L can be rewritten as

$$\begin{aligned} L &= \sum_{p=1}^{P}\Big(\sum_{i=1}^{N}\sum_{j=1}^{W_p} n_p(i,\omega_j^p)\log P(i,\omega_j^p)\Big) \\ &= \sum_{p=1}^{P}\sum_{i=1}^{N} n_p(i)\Big[\log p(i) + \dots \\ &+ \sum_{j=1}^{W_p}\frac{n_p(i,\omega_j^p)}{n_p(i)}\log\sum_{k=1}^{K} p(\omega_j^p|k)p(k|i)\Big] \end{aligned} \tag{5}$$

where $n_p(i) = \sum_{j=1}^{W_p} n_p(i,\omega_j^p)$ is the number of total features detected at the pth component in the ith video. We apply the Expectation Maximization (EM) algorithm on the objective function (5) for the learning problem. Our EM model is illustrated in Fig 2.

E-step:

$$p(k|i,\omega_j^p) = \frac{p(\omega_j^p|k)p(k|i)}{\sum_{l=1}^{K} p(\omega_j^p|l)p(l|i)} \tag{6}$$

M-step:

$$p(\omega_j^p|k) = \frac{\sum_{i=1}^{N} n_p(i,\omega_j^p)p(k|i,\omega_j^p)}{\sum_{r=1}^{W_p}\sum_{i=1}^{N} n_p(i,\omega_r^p)p(k|i,\omega_r^p)} \tag{7}$$

$$p(k|i) = \sum_{p=1}^{P}\sum_{j=1}^{W_p}\frac{n_p(i,\omega_j^p)p(k|i,\omega_j^p)}{\sum_{p=1}^{P} n_p(i)} \tag{8}$$

According to TUC, for those pairs $\{(i,\omega_{j'}^p)|\forall p \in (1\dots P)\}$ observed in the same frame, the class of motions found in them should be unique, though the component label p involved can be different. In this sense, we use TUC, not a particular precision value, to control EM iteration times, i.e., when (7) has the same class k for all visual words from the same frame, the iteration of EM stops.

B. Automatic Label Tree

Spatio-temporal features are often used to represent video events in recent research of motion understanding. Laptev et al. [17] presented a STIP detector developed from the Harris and Föstner interest point operator. Their approach can detect those points where the local image values have significant variations in spatio-temporal space. Dollár et al. [6] proposed an alternative approach to detect STIP based on separable linear filters for face expressions and mouse motion analysis. However, whether these STIP can give a compact representation for the sematic meaning of video depends on the selection of detection scales, as different video events may have different space and time extents. Hence, using a fixed scale to detect all videos, fake STIP occur. The upper row of Fig 4 shows an example of fake STIP under fixed detection scales. Laptev et al. [17] proposed an adaptive method of scale selection which can adaptively choose appropriate scales for the current detection. Their method can catch meaningful STIP for a particular video, but its application is limited to the simple and short videos. For this purpose and for the purpose of collection of the labeled visual words in our EM model, we propose an automatic label machine, that is automatic label tree (ALT), which can not only find the most meaningful STIP from the raw ones detected by fixed scales, but can also label them with their corresponding component indices. We follow Dollár et al.' method [6] to detect STIP under a fix pair of spatio-temporal scales, and represent them with gradient descriptor vector called raw GRAD descriptor [6] in our context.

As K-Means coding is difficult to capture the diversity and richness of high-dimensional descriptors. Some recent approaches [14] [23] [24] have focused on building more discriminative code books for image classification. In [21] [22], classification trees were used to classify descriptors of image by voting for the tree-assigned class labels. Moosman et al.' method [14] developed Extremely Randomized Clustering Forests (ERC-Forests) based on Extremely Randomized Trees

[16] for building visual code books of image feature descriptor SIFT [17]. Their approach is robust to background clutters and is quicker and more discriminative than K-Means method.

To eliminate those fake spatio-temporal interest points detected at the background caused by the fixed detection scales, we extend [12] to 3D field for motion clustering, but our approach works quite differently. Following tree growing algorithm in [12], we use the raw GRAD descriptors that have body component labels to guide the tree construction, tree structure is illustrated in Fig 3. When a tree has grown up, its leaf represents a set of the descriptors with the same component label, i.e. a leaf corresponds to a component, while a component can have several leaves. Hence, all leaves under the same component p can be used to build their common vocabulary $\mathbf{V}_p$, with each word recording the leaf index only, not all descriptor vectors contained in it. During a query, each local raw GRAD descriptor sampled from the test video traverses the tree from the root down to a leaf. Each query descriptor can only fall into a leaf once, if there are m total query descriptors of video i falling into the leaf ω_j^p, we say the visual word ω_j^p appears m times in video i, i.e., $n_p(i, \omega_j^p)$ is equal to m. Again, superscript p is the component label of leaf ω_j^p, consequently, those m raw descriptors can be localized according to this superscript. Or we say that one can get to know the location of descriptor in human body after a traverse, and eliminate the descriptors that fall into the leaves with background label, which are in fact corresponding to the fake STIP. Fig 4 shows the labeling performance of ALT.

Since it can automatically assign a test STIP whose component label is unknown to its corresponding component, we call it automatic label tree (ALT). As it is illustrated in Fig 3, ALT is such a random tree whose branch records an attribute index f_i of GRAD vector and its threshold θ_i, but not the whole set of attributes. In order to tell which visual word a detected GRAD vector may belong to, one just need to make comparison between the f_ith attribute value of GRAD and the threshold θ_i. Times of comparison are depending on the length of the traverse path. In this sense, ALT method avoids the heavy distance computation in high dimensional space which is usually done in K-Means approach, while still being able to yield more discriminative results. Briefly speaking, ALT is like the combination of P component coding books, i.e., using STIP extracted from the test video set $\mathbf{D}$ to traverse ALT, one can get P co-occurrence matrices $\mathbf{M}_p, p \in (1 \dots P)$ simultaneously.

III. Experiments

We test our model using KTH dataset [1]. It contains 6 motion classes performed by 25 different subjects, some example frames are shown in Fig 4. We select half of them as training samples to grow ALT, and the remaining ones as test videos. Giving a set of test videos D, one can build multiple co-occurrence matrices through ALT, and use them as input of M-PLSA. According to the output $p(k|i)$ of M-PLSA, videos in D can be categorized. Classification accuracy of our method is given in Table 1 with results of Fei-Fei et al [7] for comparison.

TABLE I
ACCURACY COMPARISON

methods	recognition accuracy(%)	learning
Our methods	83.30	unlabeled
FeiFei et al. [7]	81.50	unlabeled

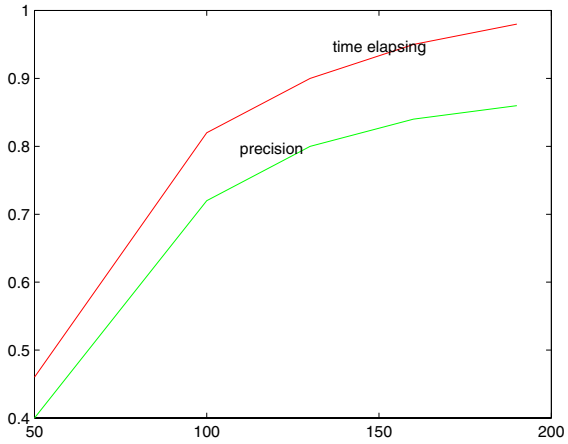


Fig. 5. Performance of M-PLSA tested on dataset KTH [1]: time elapsing has been normalized.

Advantage of our approach lies in that we categorize all input videos simultaneously rather than identify motion classes individually, the larger size of input video set, the more accurate results can be achieved. Giving a large video set as input, classification accuracy of our model is higher than Fei-Fei et al' [7], this is probably caused by more discriminative information contained in a larger co-occurrence matrix built from the input video set. Fig 5 shows precision and elapsing time of M-PLSA, conditioned on the size of test video set, i.e., classification accuracy is increasing along with the increase of the input video number. Confusion matrix of our approach is shown in Fig 6, again, we cite the results of Fei-Fei et al' [7] for comparison.

Last but not the least, since ALT can automatically label the raw STIP, it can be used to segment body components. Fig 4 reflects its performance of segmentation, from which one can find ALT can exactly label STIP, or segment components. Note that different colors for different components in Fig 4(b) are dotted automatically.

IV. Conclusion

In this paper, we presented a topic model based on multiple "bags of features" for human motion classification. Our model counts less on the body components of minor importance for motion classification. It can classify all input videos at

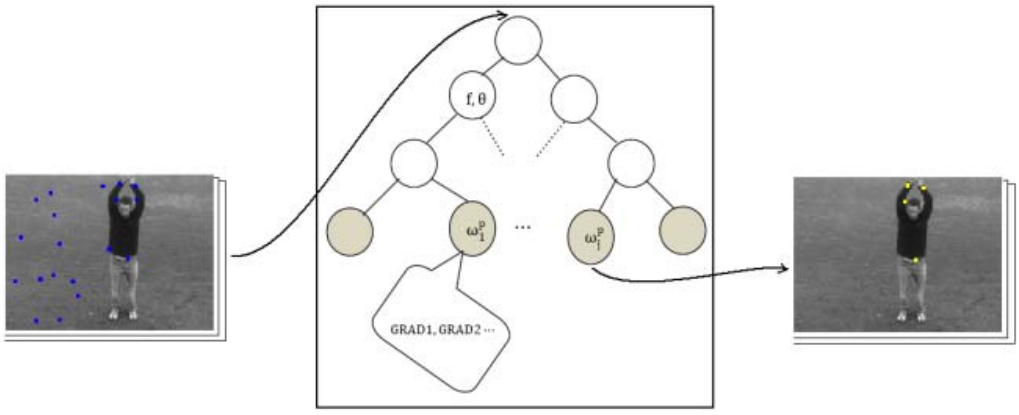
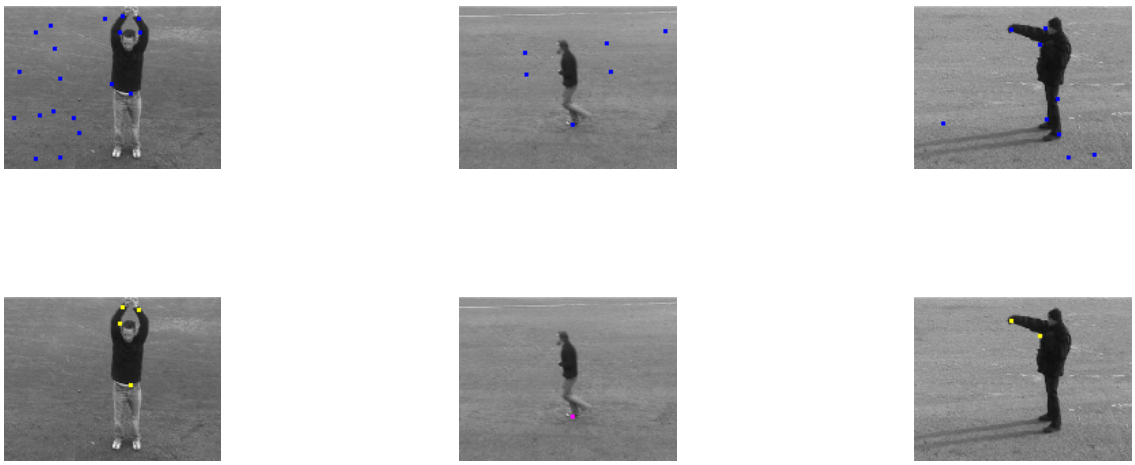


Fig. 3. Using ALT as a combination of component vocabularies in M-PLSA classification. The non-leaf node of ALT only stores one attribute index f_i of GRAD and its comparision threshold value θ_i, while the leaf stands for a visual word.



(a) waving (b) jogging (c) boxing

Fig. 4. Examples of STIP detected from KTH data set. The upper row are those unlabeled STIP detected under fixed spatio-temporal scales, they are represented by blue points and used as ALT input, and notice that there are some fake STIP caused by the giving fixed detecting scales. The lower row are their corresponding ALT output, i.e. the labeled results of query STIP shown in the upper row. Different colors indicating different components: pinks is associated to foot component while yellows indicating hands, and these filtered output are colored automatically in ALT.

the same time and in an unsupervised way. Besides, to take geometry constraints of different body component into our model, we presented an automatic label tree, which can be useful in component segmentation too.

Moreover, according to [15], random forests can beat over a decision tree in the aspects of stability and accuracy of classification, so it would be of great interests to apply random forests to do the label task in our future work.

REFERENCES

[1] C. Schúldt, I. Laptev, and B. Caputo, *Recognizing human actions: a local SVM appraoch*. In ICPR, pages 3: 32-36, 2004.

[2] M. Blank, L. Gorelick, E. Shechtman, M. Irani and R. Basri., *Actions as space-time shapes*. In ICCV, 2005.

[3] Grégory Rogez, Carlos Orrite-Orunuela and Jesús Martínez-del-Rincón, *A spatio-temporal 2D models framework for human pose recovery in minocular sequences*. Pattern Recognition, 41(9): 2926-2944, 2008.

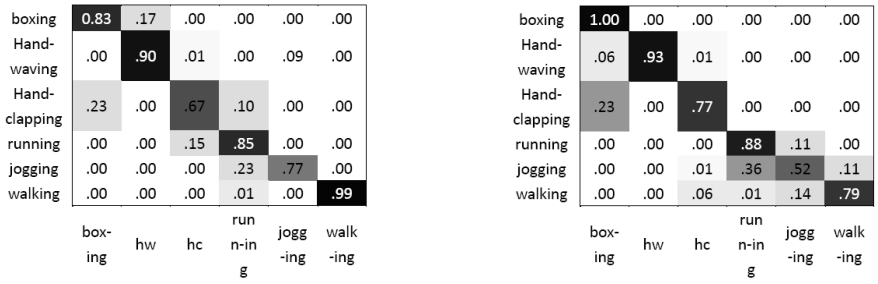
[4] Kenji Shoji, Atsushi Mito and Fubito Toyama, *Pose estimation of a 2D Articulated object from its silhouette using a GA*. In ICPR, 2000.

[5] Guo. Y., Xu. G. and Tsuji. S., *Uderstanding human motion patterns*. In ICPR, 1994.

[6] Piotr Dollár, Vincent Rabaud, Garrison Cottrell and Serge Belongie, *Behavior recognition via sparse spatio-temporal features*. In ICCV, 2004.

[7] Juan Carlos Niebles, Hongcheng Wang, and Li Fei-Fei, *Unsupervised learning of human action categories using spatial-temporal words*. International Journal of Computer Vision, 79(3): 299-318, 2008.

[8] Liu D. and Tsuhan Chen, *Unsupervised image categorization and object localization using topic models and correspondences between images*. In



(a) (b)

Fig. 6. Confusion matrices: rows are ground truth, and columns are predicted motion classes. (a) Confusion matrices of M-PLSA, and (b) is the one of Fei-Fei et al' [7].

ICCV, 2007.

[9] G. Csurka, C. Dance, L. Fan, J. Williamowski and C. Bray, *Visual categorization with bags of keypoints*. Proc. ECCV Workshop Statistical Learning in Computer Vision: 59-74, 2004.

[10] J. Sivic and A. Zisserman, *Video google: a text retrieval approach to object matching in videos*. Proc. Eighth IEEE Int'lConf. Computer Vision, vol. 2: 1470-1477, 2003.

[11] Jitendra Malik, Serge Belongie, Jianbo Shi and Thomas Leung, *Textons, contours and regions: cue integration in imagesegmentation*. In ICCV, 1999(2): 918-925, 1999.

[12] Frank Moosmann, Eric Nowak and Frederic Jurie, *Randomized clustering forests for image classification*. IEEE Transections on Pattern Analysis and Machine Intelligence, 2008.

[13] V. Lepetit, P. Lagger and P. Fua, *Randomized trees for real-time keypoint recognition*. Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition, (2): 775-781, 2005.

[14] R. Marée, P. Geurts, J. Piater and L. Wehenkel, *Random subwindows for robust image classification*. Proc. IEEE Intl Conf. Computer Vision and Pattern Recognition,(1): 34-40, 2005.

[15] Breiman L., *Random forests*. Machine Learning, (45): 5C32.

[16] P. Geurts, D. Ernst and L. Wehenke, *Extremely randomized trees*. Machile Learning Journal, 63(1), 2005.

[17] I. Laptev, *On space-time interest points*. in International Journal of Computer Vision, 64(2/3): 107-123, 2005.

[18] Jingen Liu and Mubarak Shah, *Learning human actions via information maximization*. In CVPR, 2008.

[19] Yaser Sheikh, Mumtaz Sheikh and Mubarak Shah, *Exploring the space of a human action*. In ICCV, 2005.

[20] Thomas Hofmann, *Unsupervised learning by probabilistic latent semantic analysis* . In Machine Learning, 42(1):177-196, 2001.

[21] H. Blockeel, L. De Raedt, and J. Ramon, *Top-down induction of clustering trees* . Proc. 15th Intl Conf. Machine Learning: 55-63, 1998.

[22] B. Liu, Y. Xia and P.S. Yu, *Clustering through decision tree construction*. Proc. Ninth ACM Intl Conf. Information and Knowledge Management: 20-29, 2000.

[23] D. Nistér and H. Stewénius, *Scalable recognition with a vocabulary tree*. In CVPR, 2006.

[24] J. Winn and A. Criminisi, *Object class recognition at a glance*. In CVPR06 - video tracks, 2006.