Learning A Joint Discriminative-Generative Model for Action Recognition

Ioannis Alexiou[1], Tao Xiang[2] and Shaogang Gong[1]
[1]School of Electronic Engineering and Computer Science, Queen Mary University of London, UK
[2]Vision Semantics Ltd, UK
Email: {i.alexiou, s.gong}@qmul.ac.uk, txiang@visionsemantics.com

***Abstract*—An action consists of a sequence of instantaneous motion patterns whose temporal ordering contains critical information especially for distinguishing fine-grained action categories. However, existing action recognition methods are dominated by discriminative classifiers such as kernel machines or metric learning with Bag-of-Words (BoW) action representations. They ignore the temporal structures of actions in exchange for robustness against noise. Although such temporal structures can be modelled explicitly using dynamic generative models such as Hidden Markov Models (HMMs), these generative models are designed to maximise the likelihood of the data therefore providing no guarantee on suitability for discrimination required by action recognition. In this work, a novel approach is proposed to explore the best of both worlds by discriminatively learning a generative action model. Specifically, our approach is based on discriminative Fisher kernel learning which learns a dynamic generative model so that the distance between the log-likelihood gradients induced by two actions of the same class is minimised. We demonstrate the advantages of the proposed model over the state-of-the-art action recognition methods using two challenging benchmark datasets of complex actions.**

***Keywords*—Joint Learning; Discriminative-Generative Models; HMM; FKL**

I. Introduction

Most existing work on action recognition deploys discriminative models such as kernel machines and metric learning [1], [2]. Spatio-temporal features are first extracted from video frames of action sequences; action descriptors are then constructed by discretising all the features of the entire action sequences into a Bag-of-Words (BoW) representation in the form of fixed-length histogram vectors. Taking these BoW representations of action sequences, a discriminative model utilises the given class labels of some training sequence data to learn the optimal decision boundaries for separating different action classes observed in video either in the visual spatio-temporal feature space or a kernel embedding space. In this discriminative classifier based approach, although the local temporal information is captured implicitly by the spatio-temporal features, the global temporal structure of an action is completely ignored.

However, for recognising actions of complex (and subtle) temporal structures, action recognition using a discriminative classifier model trained on a BoW representation is inadequate. Different approaches are necessary for modelling explicitly the temporal structures. To this end, there have been attempts on exploring dynamic generative models. These generative model are designed to learn explicitly temporal ordering dependencies of a sequence of data. Most of these models are probabilistic graphic models, including Dynamic Bayesian Networks (DBNs) such as Hidden Markov Models (HMM) [3], [4], probabilistic topic models (PTMs) such as Latent Dirichlet Allocation (LDA) [5] and its extensions [6], [7], or a hybrid of both DBN and PTM [8]. However, most existing generative models are constructed for abnormal behaviour detection or temporal segmentation. When deployed for action recognition, these generative models are often suboptimal. This is because such models are designed to maximise the log-likelihood of the training data required for anomaly detection or segmentation. This does not necessarily guarantee that these models are also good for discriminating actions of different classes.

In this work, we propose a new method for action recognition that explores both discriminative feature learning and generative temporal modelling, that is, to discriminatingly learn a dynamic generative model which captures explicitly action temporal structures and simultaneously optimised directly for distinguishing different action classes. Our approach is based on learning discriminative Fisher kernel using a HMM. In contrast to previous methods that learn a HMM for each class independently for maximum data likelihood, we learn all class-specific HMMs jointly and discriminatively so that the class similarity distance between the log-likelihood gradients induced by two actions of the same class is minimised whist those of different classes are made farther apart. The Fisher kernels computed using these discriminatively trained HMMs are then utilised by a kernel machine for classification.

II. Related Work

The limitation of the BoW+discriminative classier based approach for complex action recognition has long been recognised and various solutions have been proposed to overcome this limitation. One solution is to model explicitly the temporal structures of each action category. Most of such models are generative models, e.g. Dynamic Bayesian Networks (DBNs) [3], Propagation Net [9] among which graphical models especially DBNs are popular [10]. Different DBN topologies have been developed for object-based decomposition and to factorise the state space and/or observation space by introducing multiple hidden state variables and observation state variables, e.g. Multi-Observation HMM (MOHMM) [11], Parallel HMM (PaHMM) [12]. representation to discriminate complex human activities. A recent study by Kuehne et al. [13] shows that additional granularity in the action length improves the HMM performance. They demonstrate that classifying smaller components of action sequences by HMM similar to a fine-grained object classification approach can benefit action recognition. However, these generative model based methods are based on maximum likelihood learning of different class models independently for individual action classes. In contrast, our method learns HMMs for different classes jointly and discriminatively so to maximise inter-class discrimination.

III. Representation and Modelling

Spatio-temporal features are first extracted for capturing temporal structural information. In our approach, we adopt the MBH (Motion Boundary Histogram) features [14]. Local optical flows are computed and dense trajectories are formed across each entire action video sequence. MBH features are constructed along those dense trajectories across each action video. We then divide each action video into a discrete number of fixed length short clips with some overlapping between clips. In each short clip, we compute k-means clustering of MBH features and construct a dictionary of words for a BoW histogram representation of MBH features for this clip.

Formally, MBH descriptors (features) d_i are extracted along the dense trajectories in each action video $L_i = (x_i, y_i, t_i)$. A set (pool) of N MBH descriptors (features) is given as $\mathbf{P}_v = \{(d_1, L_1), \ldots, (d_i, L_i)\ \ , \ldots, (d_N, L_N)\}$ for video v, where the total number of action videos is V. To construct temporally ordered and localised BoW representations of MBH features, each action video is split into short clips with fixed-length $T \forall t$ at time t, where each clip is $\mathbf{P}_{vt} \subset \mathbf{P}_v$ and $\mathbf{P}_{vt} = \{d_1, \ldots, d_i, \ldots, d_T\}$. Each clip $\mathbf{P}_{vt}$ may overlap with another $\mathbf{P}_{vt'}$. Suppose the overlapping is between (t, T), the overlapped video segment between two clips is then $B = 1 - \frac{t'-t}{T} \iff t' - t \leq T$. In each clip, MBH features are clustered with k-means to generate K centres for constructing a localised MBH dictionary, denoted as $\mathbf{w} = \{w_1, \ldots, w_K\}$. More precisely, MBH features extracted from each action video are represented by a sequence of localised BoW representations as follows:

$$\mathbf{C}_i = \underset{k}{\text{argmin}} \sqrt{\sum_{j=1}^{J=192} (\mathbf{w}_{k_l} - \mathbf{d}_{i_l})^2} \tag{1}$$

$$f(\mathbf{C}_i, k) = \begin{cases} 1 & \text{if } \mathbf{C}_i = k \\ 0 & \text{if } \mathbf{C}_i \neq k \end{cases}$$

$$\mathbf{H}_t = \{\textstyle\sum_{n=1}^{N} f(\mathbf{C}_i, 1), \ldots, \sum_{n=1}^{N} f(\mathbf{C}_i, k), \ldots, \sum_{n=1}^{N} f(\mathbf{C}_i, K)\} \tag{2}$$

In (1) the MBH descriptors/features in each temporally segmented video clip $\mathbf{P}_{vt}$ are assigned to computed K cluster centres $\mathbf{w}$. Each MBH descriptor is assigned to the nearest centre with a cluster ID $\mathbf{C}_i$. The resulting counts of MBHs in each cluster forms a histogram $\mathbf{H}_t$, defined in (2).

A. HMM for Temporal Modelling

The clip-wise MBH histograms are fed into a HMM model to learn the temporal structures of actions. The role of HMM is to learn the temporal dependencies between different clips of an action video. More precisely, the model parameters of a HMM are trained from a set of action videos per action class, of which each action video sequence is represented by temporally segmented MBH histograms, defined by $D_v = \{H_1, \ldots, H_T\}$. Suppose the number of HMM states is given as S. This corresponds to the number of discrete video clips into which each action video is segmented[1]. The HMM model is initialised by k-means clustering that estimates a Gaussian Mixture Model (GMM) of S Gaussians $\mathbf{\Lambda} = \{\boldsymbol{w}_s, \boldsymbol{\mu}_s, \mathbf{\Sigma}_s, s = 1, \ldots, S\}$ where the mixture weights, means and covariance of each state are estimated respectively. The state dependencies are described by the transitions matrix. Given some training action sequences, these HMM parameters are estimated by the Forward-Backward algorithm through maximising the following log-likelihood:

$$\mathcal{L}(D) = \sum_{v=1}^{V} \log p(\mathbf{D}_v, \mathbf{z} | \mathbf{\Theta}) \tag{3}$$

where

$$p(\mathbf{D}_v, \mathbf{z} | \mathbf{\Theta}) = p(z_1 | \pi) \left[\textstyle\prod_{n=2}^{N} p(z_n | z_{n-1}, \mathbf{A}) \right] \textstyle\prod_{n=1}^{N} p(x_n | z_n, \beta) \tag{4}$$

The equation (4) above describes the joint probability distribution of the data $\mathbf{D}_v$ given the hidden states $\mathbf{z} = \{z_1, \ldots, z_S\}$ and model parameters $\mathbf{\Theta} = \{\pi, \mathbf{A}, \beta\}$, where the model parameters $\mathbf{\Theta}$ consists of the probabilities of the states π, transitions $\mathbf{A}$ and emissions β, where transitions and emission probabilities are used to estimate the ordering of the S states. This optimisation problem can be solved by either the variational expectation maximisation algorithm or gradient decent. Model training is performed for each action class where for simplicity, class labels have been omitted. Now, to make the learned HMMs for different action classes discriminative, we explore discriminative Fisher kernel learning as follows.

B. Discriminative Fisher Kernel Learning

Fisher kernel learning aims to enhance the discriminative properties of a model in its parameter space [15]. In this work, the Θ and Λ parameters are considered for learning discriminative Fisher kernels. For simplicity we denote $M = \{\Theta, \Lambda\}$ and m being a parameter vector for each HMM state. The following equation is differentiated:

$$O(\boldsymbol{M}, \boldsymbol{W}; \boldsymbol{D}) = \sum_{i} \sum_{j \neq i} \delta_{y_i, y_j} p_{ij} \tag{5}$$

where $\boldsymbol{W}$ is a stochastic selection matrix, δ is the Kronecker delta, and p_{ij} is the probability of data sample D_v assigning

[1] We set $S = 5$ in this work unless otherwise stated.

its label to another data sample. The probability p_{ij} in (7) are gradients computed from the data as follows:

$$\mathbf{g}_s = \frac{\partial \mathcal{L}(\mathbf{D})}{\partial m} \qquad \forall m \in M \tag{6}$$

In (6), the gradient $\mathbf{g}_s$ is computed by associating the highest posterior for each parameter vector, that is, the gradient which is computed between the actual data and the parameter vector. In practice, the HMM gradient is estimated based on the temporally aligned histograms as follows:

$$p_{ij} = \frac{exp(-(g_i - g_j)^T W^T W (g_i - g_j))}{\sum_{j' \neq i} exp(-(g_i - g_{j'})^T W^T W (g_i - g_{j'}))} \tag{7}$$

The Fisher kernel learning representations are obtained by evaluating the posterior probabilities in (4) and maximising the partial derivatives of (5) with $(\frac{\partial O}{\partial m}, \frac{\partial O}{\partial W})$, where

$$\frac{\partial O}{\partial m} = 2Q_m \sum_i \Bigg[\sum_{j \neq i} \Big[\delta_{y_i, y_j} p_{ij} (\mathbf{g}_i - \mathbf{g}_j) \Big(\frac{\partial \mathbf{g}_i}{\partial m} - \frac{\partial \mathbf{g}_i}{\partial m} \Big) \Big] - p_i \sum_{j \neq i} \Big[p_{ij} (\mathbf{g}_i - \mathbf{g}_j) \Big(\frac{\partial \mathbf{g}_i}{\partial m} - \frac{\partial \mathbf{g}_i}{\partial m} \Big) \Big] \Bigg] \tag{8}$$

Q_m represents an element of the diagonal matrix $Q = \mathbf{W}^T \mathbf{W}$ that corresponds to the model parameter m and $p_i = \sum_j p_{ij}$. Therefore,

$$\frac{\partial O}{\partial \mathbf{W}} = 2\mathbf{W} \ \frac{\partial O}{\partial Q} \tag{9}$$

$$\tfrac{\partial O}{\partial Q} = - \textstyle\sum_i \sum_{j \neq i} \delta_{y_i, y_j} p_{ij} \| \mathbf{g}_i - \mathbf{g}_j \|^2 - p_i \sum_{j \neq i} p_{ij} \| \mathbf{g}_i - \mathbf{g}_j \|^2 \tag{10}$$

The final Fisher kernel is estimated by the following equation:

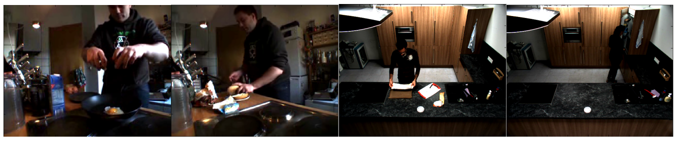
$$k(\mathbf{D}_{vi}, \mathbf{D}_{vj}) = \mathbf{g}_i^T Q \ \mathbf{g}_j \tag{11}$$

That is, the final kernel is computed as a dot product between the discriminant gradients and the diagonal matrix Q. After obtaining the final Fisher kernel representation, we use the kernel matrix in a standard SVM to obtain a classifier. In this way the generative model (HMM) and discriminative classifier (SVM) are seamlessly combined by the Fisher kernel learning. We call this new model for action recognition the Fisher Kernel Learning of Hidden Markov Model (FKL-HMM).

IV. Experiments

The proposed FKL-HMM model was evaluated against (1) the standard HMM method and two state-of-the-art action recognition model using (2) a BoW representation with SVM [2] and (3) a structured temporal process with hierarchical HMM (HTK) [13]. The experiments were carried out on two challenging action datasets, the Breakfast [13] and Cooking activities [2]. Some examples are shown in Figure 1.

For the BoW representation model of [2] the MBH features [16], [14] are used. Each action video is then represented individually by a 4K dimensional histogram (holistic codebook) of MBH features where are trained and tested using a multi-label χ^2-SVM . Note, in this study we focus on analysing the role and importance of modelling temporal structures in action recognition. Therefore no appearance based descriptors are used in the experiments. For constructing the proposed FKL-HMM model, temporally localised MBH features are encoded into hisotgrams of fixed-length short video length for each action video, where each has 80 frames with 75% overlap between neighbouring histograms in a video. A codebook of 300 words is used for the histogram encoding. These temporal sequences of histograms are trained with an HMM model (Sec. III-A) and the proposed new FKL-HMM model (Sec. III-B). On comparative evaluation, Multi-class Classification Accuracy (MCA) is adopted in this study for benchmarking both datasets. The MCA criterion is selected as a harder metric than Mean Average Precision (mAP) because MCA evaluates the top ranked results comparing to mAP which gives a softer average of the ranking. In addition, a Relative Performance Gain (RPG) measurement is also reported on both datasets. This measure is established by subtracting the recognition accuracy per class between two comparing methods and recording the mean performance margins respectively. For training and testing of all the models, we follow the training and testing data splits as defined in [13] and [2] respectively.



(a) FriedEgg (b) Sandwich (c) UnrollDough (d) TakeOutFromFridge

Figure 1. Action instances from both the Breakfast ((a) & (b)) and the Cooking ((c) & (d)) datasets. Each of the two datasets contains both short/simple ((b) & (d)) and long/complex ((a) & (c)) actions in duration respectively.

TABLE I

Action recognition comparisons on the Breakfast and the Cooking Datasets.

	FKL-HMM	HMM	BoW[2]	HTK [13]	BoW(50CB)[13]
Breakfast [13]	52.01	49.53	—	40.50	26.00
Cooking [2]	41.06	35.21	40.50	—	—

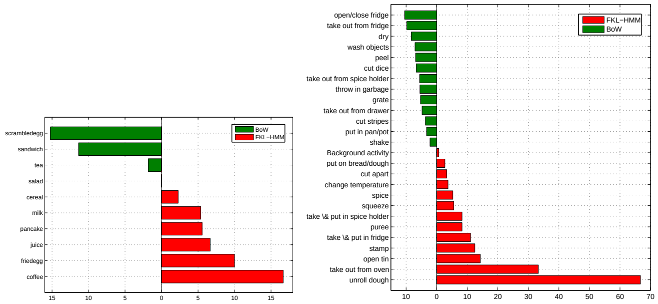
Table I shows the comparisons of action/activity recognition performance of different models, including FKL-HMM, HMM, BoW [2], HTK [13] and BoW(50CB) [13][2], on both the Breakfast and the Cooking datasets. The reported performance measure is Multi-class Classification Accuracy (MCA). It is evident that the proposed new FKL-HMM model outperforms all other models. The structured temporal model HTK has clear advantage over the BoW holistic model (BoW(50CB)) with the same features for the recognition of more complex actions/activities with longer durations as in the case of the

[2] This differs from BoW [2] in both the features (HOG/HOF vs. MBH) and code book (CB) size (50 vs. 300).

Breakfast dataset. Moreover, the Cooking dataset is annotated at a finer level of actions of cooking activities. This resulted in difficulties in distinguishing many short actions. Our results show that the Fisher kernel learning in the HMM parameter space provides notable advantages over the HMM model alone, therefore better suited for fine-grained action discrimination. The performance using only the MBH features is reported to allow fair comparisons with their benchmarks.

A more detailed analysis on different models' effectiveness on different types of action/activities in terms of short/long and simple/complex temporal structures, is shown with comparative evaluations on the Relative Performance Gain (RPG) between a BoW holistic representation based model and the proposed FKL-HMM model. For the Breakfast dataset, it is evident in Figure 2a that the FKL-HMM model performs better than BoW at six categories of activities whilst the latter is better at three, and both are equally good at one category. Among the six activity classes where FKL-HMM performs better, the average action temporal durations are significantly longer than those of the three classes when BoW do better. For example, FKL-HMM has a 10% RPG on "Fried egg" which has an averaged duration of 3,180 frames, whilst BoW has a 12% RPG on "Sandwich" whose average duration is 1,575 frames.

For the Cooking dataset, it is evident in Figure 2b that the learned temporal ordering information by the FKL-HMM model is very important for some classes but not for all of them. In particular, actions/activities such as "unroll dough", "take out from oven" and "open tin" are better recognised by the FKL-HMM model whilst actions such as "open/close fridge", "take out from fridge", and "dry" are better recognised by the BoW model. By observing these actions, one notices clearly that actions with short temporal durations (and repetitions) are better recognised by the BoW model. In contrast, those actions with more variability and longer temporal durations are better recognised by the FKL-HMM model. In particular, it is noted that for the 10 classes that our FKL-HMM has the biggest RPGs, their average duration is 416 frames, whilst for the 10 classes whereby BoW has the biggest winning margin, the average duration is 156 frames.



(a) Breakfast

(b) Cooking

Figure 2. The Relative Performance Gain (RPG) between the FKL-HMM and the BoW models on both the Breakfast and the Cooking datasets. The horizontal axis shows the per action/activity class RPG and the vertical axis denotes the action/activity class labels.

V. Conclusion

In this work we introduced a novel action recognition model by exploring Fisher kernel learning in the Hidden Markov Model parameter space that represents the temporal structures of action sequences. Our proposed FKL-HMM model aims to learn a discriminative function in a generative temporal model parameter space, therefore being benefited from both the strengths of discriminating learning and generative modelling necessary for effective recognition of more complex and longer-duration actions and activities in more realistic environments. Our model is evaluated comparatively against existing and the state-of-the-arts structured temporal models and the BoW holistic classification models using two challenging action benchmark datasets, the Cooking and the Breakfast datasets. Our experiments demonstrate clearly that the proposed new FKL-HMM model is advantageous over the existing state-of-the-arts models using either BoW holistic representations or structured temporal modelling. Future work involves the deployment of active and transfer leaning methods to explore further improvement on the model.

Acknowledgements

The research leading to these results has received funding from the European Union's Seventh Framework Programme managed by REA-Research Executive Agency http://ec.europa.eu/research/rea ([FP7/2007-2013] [FP7/2007-2011]) under grant agreement No 606952- SmartPrevent.

References

[1] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," *CoRR*, 2012.

[2] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, "A database for fine grained activity detection of cooking activities," in *CVPR*, 2012.

[3] T. Xiang and S. Gong, "Video behaviour profiling for anomaly detection," *TPAMI*, vol. 30, no. 5, pp. 893–908, 2008.

[4] C. C. Loy, T. Xiang, and S. Gong, "Detecting and discriminating behavioural anomalies," *PR*, vol. 44, no. 1, pp. 117–132, 2011.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, no. 3, pp. 993–1022, 2003.

[6] Y. Wang and G. Mori, "Human action recognition by semilatent topic models," *TPAMI*, vol. 31, no. 10, pp. 1762–1774, 2009.

[7] J. Li, S. Gong, and T. Xiang, "Learning behavioural context," *IJCV*, vol. 97, no. 3, pp. 276–304, 2012.

[8] E. Jouneau and C. Carincotte, "Particle-based tracking model for automatic anomaly detection," in *ICIP*, 2011, pp. 513–516.

[9] Y. Shi, A. Bobick, and I. Essa, "Learning temporal sequence model from partially labeled data," in *CVPR*, 2006, pp. 1631–1638.

[10] T. B. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *CVIU*, vol. 104, no. 2, pp. 90–126, 2006.

[11] S. Gong and T. Xiang, "Recognition of group activities using dynamic probabilistic networks," in *ICCV*, 2003, pp. 742–749.

[12] C. Vogler and D. Metaxas, "A framework for recognizing the simultaneous aspects of American sign language," *CVIU*, vol. 81, no. 3, pp. 358–384, 2001.

[13] H. Kuehne, A. Arslan, and T. Serre, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *CVPR*, Columbus, USA, June 2014.

[14] H. Wang and C. Schmid, "Action Recognition with Improved Trajectories," in *ICCV*, 2013.

[15] L. van der Maaten, "Learning discriminative fisher kernels," in *ICML*, 2011.

[16] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *ECCV*, 2006.