Hierarchical Bayesian Multiple Kernel Learning Based Feature Fusion for Action Recognition

Wen Sun[1], Chunfeng Yuan[1⋆], Pei Wang[1], Shuang Yang[1], Weiming Hu[1], and Zhaoquan Cai[2]

[1] CAS Center for Excellence in Brain Science and Intelligence Technology, National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
{wen.sun,cfyuan,pei.wang,syang,wmhu}@nlpr.ia.ac.cn
[2] Huizhou University, Huizhou, Guangdong, China
gd0752888@126.com

Abstract. Human action recognition is an area with increasing significance and has attracted much research attention over these years. Fusing multiple features is intuitively an appropriate way to better recognize actions in videos, as single type of features is not able to capture the visual characteristics sufficiently. However, most of the existing fusion methods used for action recognition fail to measure the contributions of different features and may not guarantee the performance improvement over the individual features. In this paper, we propose a new *Hierarchical Bayesian Multiple Kernel Learning* (HB-MKL) model to effectively fuse diverse types of features for action recognition. The model is able to adaptively evaluate the optimal weights of the base kernels constructed from different features to form a composite kernel. We evaluate the effectiveness of our method with the complementary features capturing both appearance and motion information from the videos on challenging human action datasets, and the experimental results demonstrate the potential of HB-MKL for action recognition.

Keywords: action recognition, feature fusion, multiple kernel learning

1 Introduction

Action recognition is an active research area in computer vision motivated by the promise of applications in broad domains such as intelligent surveillance, human-computer interaction and video retrieval. However, the task is still challenging due to the variations in action performances, background clutter, illumination changes, camera movements and occlusions.

The previous researches [1–7] in the literature have paid more attention to designing descriptive features which are specific to action recognition and a large number of features are available now for this task. It is an intuitive way to integrate diverse types of informative features instead of a single one to improve the

⋆ Corresponding author.

recognition performance. However, the existing action recognition algorithms [8, 9] usually employ the simple combination of different features. The most common method is the feature-level fusion which concatenates all the feature vectors together into one single feature vector. A drawback of the method is the high dimensionality of the final concatenated vector, since the efficiency of the method drops exponentially as the dimensionality increases. Another feasible solution is the kernel-level fusion. For instance, the multi-channel approach proposed in [10] simply takes the multiplication of the kernels. Nevertheless, the method cannot guarantee the performance improvement over the individual features. It is worth noting that both methods do not consider the relative importance of the candidate features and this leads to a meaningless combination. Therefore, it requires to formulate a combination method that is able to evaluate the relative contributions of different feature representations and utilize such information to gain enhanced classification performance.

In this paper, we propose a new *Hierarchical Bayesian Multiple Kernel Learning* (HB-MKL) framework to deal with feature fusion problem for action recognition. We first formulate the multiple kernel learning problem as a decision function based on a weighted linear combination of the base kernels, and then develop a hierarchical Bayesian framework with three layers to solve this problem. Specifically, the bottom layer consists of the parameters in the decision function. On the middle layer, the priors of Gaussian distribution family are placed on the parameters of the decision function. Especially, the prior on the kernel weight is set by a half-normal distribution, which has the advantage of interpretability due to the only nonnegative restriction in nature. The top layer is composed of the hyper-priors, invoked on the parameters of the priors at the level below. Gamma distribution is employed to take the advantage of the conjugacy and non-informativeness. The non-informativeness ensures that the learnt model parameters are intrinsic to the data. The model is established in a fully conjugate manner, offering the probability of efficient inference. Therefore, we derive a variational approximation for inference. After evaluating the optimal weights of the base kernels using the framework above, we derive the composite kernel. Finally, an SVM classifier is trained using the learnt optimal combined kernel. We apply the above model to the feature fusion problem in the field of action recognition, where no such attempts have been made before to the best of our knowledge. We conduct a set of experiments for better illustration and comparison on several public action datasets. The experimental results demonstrate the effectiveness of our method and provide some insight on the contributions of different features for action recognition.

The main contributions of this work can be summarized as follows. First, a new framework of hierarchical Bayesian multiple kernel learning is designed. The half-normal distribution prior placed on the base kernel weights makes them nonnegative without any other constraints, which exactly meets the actual requirements and has good interpretability. Second, instead of conventional simple fusion of multiple features used in action recognition, we propose to apply the HB-MKL based feature fusion method to action recognition, which can learn the

optimal combination of multiple features automatically. Third, we carry out a set of experiments on three datasets, and the experimental results demonstrate the efficiency of the proposed method. It is worth mentioning that the valuable results of the feature weights learnt by our method give some insight on how each feature contributes to recognizing an action.

2 Related Work

In this section, we give a brief overview of the related work on three aspects: discriminative features for action recognition, feature fusion methods and multiple kernel learning algorithms.

Various classical video feature descriptors are proposed in previous work including HOG, HOF [1], MBH [2] and some spatio-temporal extensions of image descriptors, such as 3D-SIFT [3], HOG3D [4] and extended SURF [5]. Moreover, trajectory features are also popular descriptors. In [6], human actions are represented using sparse SIFT-based trajectory. Wang *et al.* [7] introduce an approach to combine dense sampling with feature tracking, and extract robust features along the trajectories.

Realizing it is not enough to describe videos using homogeneous descriptor, some researchers try to fuse heterogeneous descriptors to construct more discriminative classifiers. However, most of the existing algorithms combine multiple features in an easy way. Tian *et al.* [8] combine the histogram of MHI and Haar wavelet transform of MHI at the feature-level. They use the straightforward concatenation of the features as the combined feature representation, which is a higher dimensional vector. Ullah *et al.* [9] use a multi-channel approach proposed in [10] to integrate feature representations, which takes the multiplication of the feature kernels in nature. The method can be regarded as a combination at the kernel-level using fixed rules without additional parameters. However, the above mentioned methods do not take into account the contribution of different features and hence cannot make better use of the multiple features. In this paper, we employ Multiple Kernel Learning (MKL) to informatively combine diverse features for action recognition.

Many variants of MKL have been proposed in the previous work. In this paper, we consider MKL with a weighted linear combination of the base kernels under a Bayesian framework. The existing Bayesian MKL methods differ in the prior assumptions on the kernel weights. Girolami *et al.* [11] present a Bayesian model for regression and classification problems by employing a Dirichlet prior on the kernel weighting coefficients. Damoulas *et al.* [12] use a similar model with the same prior distribution assumptions and extend the model for multi-class problem. Moreover, they apply the approach to protein fold recognition and remote homology detection problems to prove the validity of the method. Gönen [13] presents an efficient MKL algorithm by assuming the kernel weights to be normally distributed. In this paper, we introduce a half-normal distribution on the kernel weights. Compared with the normal distribution prior, the half-

normal distribution ensures that the kernel weights are nonnegative and hence it produces a more meaningful combination of kernels.

3 Hierarchical Bayesian Multiple Kernel Learning for Action Recognition

In this section, we first introduce the heterogeneous and complementary features used to sufficiently represent the actions in videos. Then we introduce the detailed HB-MKL algorithm and its inference. Finally, we apply HB-MKL to effectively fuse the obtained multiple features for action recognition.

3.1 Multiple Features for Action Representation

In this paper, we use the state-of-the-art improved dense trajectory features [14] for action representation. We first extract the trajectories by densely sampling feature points in each frame and tracking them in the video based on displacement information from the optical flow field. Subsequently, we compute the trajectory-aligned descriptors (i.e., Trajectory, HOF, HOG and MBH) within a space-time volume along the trajectories.

It is worth noting that the extracted features are complementary in describing action sequences by capturing both static appearance and dynamic motion information. The trajectory descriptor is a concatenation of normalized displacement vectors which describe the motion of the trajectories. HOF captures the motion information based on the orientation of optical flow, whereas HOG calculates the histograms of oriented gradients which measure the static appearance information. MBH (motion boundary histogram) encodes relative motion information by computing derivatives separately for the horizontal and vertical components of the optical flow.

Once we obtain the features above, we encode them using both Bag of Features (BOF) and Fisher Vector (FV) [15] approaches to achieve the final video sequence representations. Using one of these two strategies, each video is represented by four kinds of features which characterize complementary information of the video sequence.

3.2 Hierarchical Bayesian Multiple Kernel Learning

In order to formulate a better combination of the obtained multiple features, we propose a HB-MKL model for feature fusion. First, we formulate the MKL for multi-class classification problem as described below.

Consider N independent and identically distributed training instances $\{\mathbf{x}_i\}_{i=1}^{N}$, where each data instance has P feature representations $\mathbf{x}_i = \{\mathbf{x}_i^m\}_{m=1}^{P}$. In this paper, we consider a combined kernel which fuses different kinds of feature kernels in a linear way as follows:

$$K_e(\mathbf{x}_i, \mathbf{x}_j) = \sum_{m=1}^{P} e_m K_m(\mathbf{x}_i^m, \mathbf{x}_j^m) \,, \tag{1}$$

Table 1. List of notations

Notations	Dimensions	Representations
$\{\mathbf{K}_m\}_{m=1}^P$	$N \times N$	Base kernel matrices
$\mathbf{A}$	$N \times K$	Training instance weight matrix
$\boldsymbol{\lambda}$	$N \times K$	Priors for training instance weight matrix
$\boldsymbol{e}$	P	Kernel weight vector
$\boldsymbol{\omega}$	P	Priors for kernel weight vector
$\boldsymbol{b}$	K	Bias vector
$\boldsymbol{\gamma}$	K	Priors for bias vector
$\mathbf{F}$	$K \times N$	Classification score matrix
$\mathbf{Y}$	$K \times N$	Class label matrix

where K_m is the base kernel calculating a similarity metric between videos with respect to the m-th feature, e_m is the corresponding kernel weight indicating the m-th base kernel's contribution and significance, and K_e is the composite kernel that finally measures the overall similarity between two videos. Based on the obtained composite kernel K_e, the decision function for a test instance $\mathbf{x}_*$ with respect to action class c can be written as:

$$f^c(\mathbf{x}_*) = \sum_{i=1}^{N} a_c^i K_e(\mathbf{x}_i, \mathbf{x}_*) + b_c, \quad c = 1, \cdots, K \ , \tag{2}$$

where K is the number of the action classes, a_c^i denotes the weight assigned to the i-th training instance for the c-th action class, and b_c is the bias for the c-th action class.

We then propose a hierarchical probabilistic model to solve the above multi-class multiple kernel learning problem in a Bayesian manner. Specifically, we impose that the kernel weight e_m is sampled from a half-normal distribution with precision ω_m, which ensures that the kernel weights are non-negative without any other constraints. The training instance weight a_c^i and the bias b_c are placed by two zero-mean Gaussian distributions with precisions λ_c^i and γ_c, respectively. Thus according to the decision function, the classification score f_i^c is generated from a Gaussian distribution with the mean $\boldsymbol{e}^{\mathrm{T}} \boldsymbol{a}_c^{\mathrm{T}} \boldsymbol{k}_{m,i} + b_c$ and precision 1. Given the classification score f_i^c, the corresponding class label y_i^c is simply obtained by setting a threshold ν.

Finally, three non-informative Gamma distributions with different shape and scale parameters are placed on the precisions ω_m, λ_c^i and γ_c of Gaussian distributions respectively. On one hand, the parameters of Gamma distribution are in general non-informative and thus the learnt kernel weights, training instance weights, and biases are intrinsic to the data without prior knowledge assumptions. On the other hand, the above hierarchical probabilistic model is constructed in the conjugate exponential family, and therefore inference can be implemented via variational Bayesian or Gibbs-sampling analysis, with analytic update equations. The variables mentioned above correspond to one instance with respect

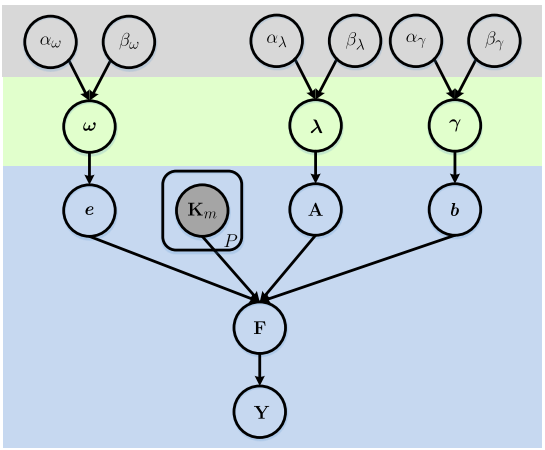

Fig. 1. Graphical model of hierarchical Bayesian multiple kernel learning

to one action class. The vector or matrix forms of these variables corresponding to all the training instances are listed in Table 1 for clarity. Actually, the superscripts and subscripts in the notations a_c^i, λ_c^i, f_i^c, y_i^c denote the row and column indexes of their matrices, respectively.

With these parametric definitions, the probabilistic graphical model of our HB-MKL framework for multi-class classification is illustrated in Fig. 1. Corresponding to the three layers in the graphical model, the proposed HB-MKL is expressed in the following three groups of formulations in summary. On the bottom layer, the classification score of the instance with respect to action class c is expressed as:

$$\begin{aligned} f_i^c|b_c, \boldsymbol{e}, \boldsymbol{a}_c, \boldsymbol{k}_{m,i} &\sim \mathcal{N}(f_i^c; \boldsymbol{e}^{\mathrm{T}} \boldsymbol{a}_c^{\mathrm{T}} \boldsymbol{k}_{m,i} + b_c, 1) \\ y_i^c|f_i^c &\sim \delta(f_i^c y_i^c > \nu) \ , \end{aligned} \tag{3}$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$, and $\delta(\cdot)$ represents the Kronecker delta function.

On the middle layer, the half-normal distribution and Gaussian distribution are placed on the parameters of the decision function, which are expressed as:

$$\begin{aligned} e_m|\omega_m &\sim \mathcal{N}^+(e_m; 0, \omega_m^{-1}) \\ a_c^i|\lambda_c^i &\sim \mathcal{N}(a_c^i; 0, (\lambda_c^i)^{-1}) \\ b_c|\gamma_c &\sim \mathcal{N}(b_c; 0, \gamma_c^{-1}) \ , \end{aligned} \tag{4}$$

where $\mathcal{N}^+(\cdot; 0, \boldsymbol{\Sigma})$ denotes a half-normal distribution with the mean vector 0 and the covariance matrix $\boldsymbol{\Sigma}$.

On the top layer, non-informative gamma hyper-priors are placed on ω_m, λ_c^i and γ_c as follows:

$$\begin{aligned} \omega_m &\sim \mathcal{G}(\omega_m; \alpha_\omega, \beta_\omega) \\ \lambda_c^i &\sim \mathcal{G}(\lambda_c^i; \alpha_\lambda, \beta_\lambda) \\ \gamma_c &\sim \mathcal{G}(\gamma_c; \alpha_\gamma, \beta_\gamma) \ , \end{aligned} \tag{5}$$

where $\mathcal{G}(\cdot; \alpha, \beta)$ denotes a Gamma distribution with the shape and scale parameters α and β.

3.3 Variational Inference

In order to perform efficient processing, we derive variational approximation methodology for inference. The variational method [16], offers a lower bound on the model evidence using an ensemble of factored posteriors to approximate the joint parameter posterior distribution. By defining the sets of priors as $\boldsymbol{\Xi} = \{\boldsymbol{\gamma}, \boldsymbol{\lambda}, \boldsymbol{\omega}\}$, hyper-priors as $\boldsymbol{\zeta} = \{\alpha_\gamma, \beta_\gamma, \alpha_\lambda, \beta_\lambda, \alpha_\omega, \beta_\omega\}$, and the remaining variables as $\boldsymbol{\Theta} = \{\mathbf{A}, \boldsymbol{b}, \boldsymbol{e}, \mathbf{F}\}$, the factorable ensemble approximation of the required posterior can be written as

$$p(\boldsymbol{\Theta}, \boldsymbol{\Xi} | \boldsymbol{\zeta}, \{\mathbf{K}_m\}_{m=1}^P, \mathbf{Y}) \approx q(\boldsymbol{\Theta}, \boldsymbol{\Xi}) = q(\boldsymbol{\lambda}) q(\mathbf{A}) q(\boldsymbol{\omega}) q(\boldsymbol{e}) q(\boldsymbol{\gamma}) q(\boldsymbol{b}) q(\mathbf{F}) \ , \tag{6}$$

and each factor in the ensemble can be defined as:

$$\begin{aligned} q(\boldsymbol{\lambda}) &= \prod_{i=1}^{N} \prod_{c=1}^{K} \mathcal{G}(\lambda_c^i; \alpha(\lambda_c^i), \beta(\lambda_c^i)) \\ q(\mathbf{A}) &= \prod_{c=1}^{K} \mathcal{N}(\boldsymbol{a}_c; \mu(\boldsymbol{a}_c), \Sigma(\boldsymbol{a}_c)) \\ q(\boldsymbol{\omega}) &= \prod_{m=1}^{P} \mathcal{G}(\omega_m; \alpha(\omega_m), \beta(\omega_m)) \\ q(\boldsymbol{e}) &= \mathcal{N}^+(\boldsymbol{e}; \mu(\boldsymbol{e}), \Sigma(\boldsymbol{e})) \\ q(\boldsymbol{\gamma}) &= \prod_{c=1}^{K} \mathcal{G}(\gamma_c; \alpha(\gamma_c), \beta(\gamma_c)) \\ q(\boldsymbol{b}) &= \mathcal{N}(\boldsymbol{b}; \mu(\boldsymbol{b}), \Sigma(\boldsymbol{b})) \\ q(\mathbf{F}) &= \prod_{c=1}^{K} \prod_{i=1}^{N} \mathcal{TN}(f_i^c; \mu(f_i^c), \Sigma(f_i^c), \rho(f_i^c)) \ . \end{aligned}$$

We can bound the model evidence using Jensen's inequality:

$$\begin{aligned} \log p(\mathbf{Y} | \boldsymbol{\zeta}, \{\mathbf{K}_m\}_{m=1}^P) \geq \\ \mathrm{E}_{q(\boldsymbol{\Theta}, \boldsymbol{\Xi})}[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\Xi} | \boldsymbol{\zeta}, \{\mathbf{K}_m\}_{m=1}^P)] - \mathrm{E}_{q(\boldsymbol{\Theta}, \boldsymbol{\Xi})}[\log q(\boldsymbol{\Theta}, \boldsymbol{\Xi})] \ , \end{aligned} \tag{7}$$

and optimize it with respect to the distribution in the following form

$$q(\tau) \propto \exp(\mathrm{E}_{q(\{\boldsymbol{\Theta}, \boldsymbol{\Xi}\} \setminus \tau)}[\log p(\mathbf{Y}, \boldsymbol{\Theta}, \boldsymbol{\Xi} | \boldsymbol{\zeta}, \{\mathbf{K}_m\}_{m=1}^P)]) \ . \tag{8}$$

3.4 HB-MKL Based Feature Fusion for Action Recognition

In order to utilize the proposed method for action recognition, we first extract and encode the features described above to get the final video descriptors. When adopting BOF representations, we use RBF-χ^2 kernel [1] to separately calculate the base kernels corresponding to different features. As for FV representations, we compute the base kernels using linear kernel function. After that, we apply the proposed HB-MKL to construct a composite kernel by learning the optimum linear combination of the multiple kernels. Finally, we train a standard SVM classifier with the combined kernel. For all the experiments, the multi-class classification is made using the one-vs-all strategy.

4 Experiments

We evaluate our method on three popular human action datasets: KTH, UCF sports, and HMDB51 datasets.

The **KTH** dataset [17] contains six action classes. The actions are performed several times by 25 subjects under 4 different scenarios. The backgrounds are relatively homogeneous and static in most sequences. We follow the experimental settings in [17] where the videos are divided into a training set (16 subjects) and a test set (9 subjects). For evaluation, the average accuracy over all classes is reported.

The **UCF sports** dataset [18] includes 150 sequences of 10 classes of human actions. The videos are extracted from sports broadcasts which are recorded in unconstrained environments with camera motion and different viewpoints. We apply a leave-one-out cross validation scheme and the evaluation is measured using the average accuracy over all classes.

The **HMDB51** dataset [19] contains a total of 6766 video clips collected from various sources, ranging from digitized movies to YouTube. The videos in the dataset vary in video quality, camera motion, viewpoints and occlusions. In our experiments, we adopt the original experimental setup as in [19] with three train/test splits. The average accuracy over the three splits is reported as the performance measurement.

4.1 Baseline Feature Fusion Methods

To evaluate the performance improvement achieved using HB-MKL, we perform experiments with two baseline feature fusion methods for comparison: concatenation and multi-channel methods. The concatenation method directly concatenates all the feature representations together to form a combined representation. The multi-channel method combines different descriptors as follows [10]:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\sum_m \frac{1}{A^m} D(\mathbf{x}_i^m, \mathbf{x}_j^m)) \,, \tag{9}$$

where $D(\mathbf{x}_i^m, \mathbf{x}_j^m)$ is the χ^2 distances defined on histogram representations between videos $\mathbf{x}_i$ and $\mathbf{x}_j$ with respect to channel m. A^m is the normalization

Table 2. Performance comparisons of five single feature based approaches as well as three fusion approaches using baseline and HB-MKL

Approaches	KTH(%)	UCF(%)	HMDB51(%)
Trajectory	92.13	82.67	33.27
HOF	94.44	85.33	40.37
HOG	87.96	84.00	28.93
MBHx	93.98	82.67	35.80
MBHy	94.44	82.67	42.16
Concatenation	93.98	78.67	39.65
Multi-channel	94.44	77.33	41.33
HB-MKL	**95.37**	**90.00**	**52.07**

factor computed as the average value of χ^2 distances between all the training instances for the m-th channel.

4.2 Comparison of Experimental Results

In order to qualify the effectiveness of our approach, we evaluate the classification accuracies achieved by each of the features alone, as well as feature combination via HB-MKL. The results of these approaches using BOF encoding are shown in Table 2. It is clear that feature fusion using HB-MKL outperforms their uses separately on all the datasets. By combining all the features using HB-MKL, we obtain 95.37% on KTH which is around 1% better than the best single feature, whereas on UCF sports it is around 5%. The improvement is even higher on HMDB51, i.e., around 10%. The results demonstrate that the integration of diverse features using HB-MKL enhances the performance compared with single feature based approach.

In addition, we also compare our method with the baseline combination methods in Table 2. It can be seen that there is a significant performance gain of our combination method over the baselines. Moreover, we notice that the combinations using baselines can not guarantee the improvement with respect to every single features. In contrast, our method consistently outperforms all single features on all the datasets. The advantage of our feature fusion method over baselines can be attributed to the ability of learning the relative importance of each feature.

We also do a performance comparison using different feature encoding strategies. Table 3 lists the results using both BOF and FV for feature encoding. We notice that the improvement of FV over BOF on the KTH dataset is slightly, whereas it reaches 4.6% on HMDB51. Unexpectedly, the performance of FV is inferior to BOF on UCF sports. Based on this evaluation, we choose the best performed FV encoding for KTH and HMDB51, and BOF encoding for UCF sports in the rest of the experiments.

We also compare our method with the most recent results reported in the literature on the three datasets in Table 4. On KTH, our method yields better

Table 3. Comparison of feature encoding strategies using BOF and FV

	BOF	FV
KTH(%)	95.37	**95.83**
UCF(%)	**90.00**	88.00
HMDB51(%)	52.07	**56.67**

Table 4. Performance comparisons of our method with the state-of-the-art results

KTH		UCF sports		HMDB51	
Sun *et al.* [23]	93.1%	Sun *et al.* [23]	86.6%	Yang *et al.* [24]	53.9%
Zhang *et al.* [21]	94.8%	Zhang *et al.* [21]	87.5%	Wu *et al.* [22]	56.4%
Veeriahet *et al.* [25]	94.0%	Lan *et al.* [26]	83.6%	Shao *et al.* [27]	49.8%
Wang *et al.* [28]	94.5%	Wang *et al.* [28]	86.7%	Liu *et al.* [29]	51.4%
Sheng *et al.* [20]	95.0%	Sheng *et al.* [20]	87.3%	Liu *et al.* [30]	48.4%
Our method	**95.8%**	**Our method**	**90.0%**	**Our method**	**56.7%**

performance than [20]. The work of [20] uses direction-dependent feature pairs to represent actions, and achieves a recognition rate of 95.0%. Zhang et al. [21] report 87.5% on UCF sports by using a simplex-based orientation decomposition descriptor to describe 3D visual features. We further improve their results by 2.5%. On HMDB51, Wu et al. [22] report 56.4% with a VLAD-based video encoding for human action recognition. We achieve 56.7% which is slightly better than theirs. It can be seen that the proposed method achieves a comparable performance to the state-of-the-art approaches.

4.3 Analysis of Feature Weights Learnt by HB-MKL

Table 5 shows the feature weights learnt by HB-MKL in the range [0, 1]. From the table, we can see how each feature contributes to the final decision. It is clearly to see that on KTH, among all the feature representations, HOF plays the dominant role, while HOG tends to have the lowest weight. This reveals that motion-based features of a video are the most informative features for action recognition on KTH. This may be because the variation in appearances between frames is very small on KTH.

As for UCF sports and HMDB51, it can be seen that HOG ranks first, followed by motion-based features. This is probably because both of the datasets contain lots of camera motion which reduces the reliability of motion-based features. Moreover, the UCF sports dataset often involves specific environment and equipment, and hence the appearance-based feature is more important for it. Therefore, it demonstrates that the proposed HB-MKL is able to learn the optimal feature weights from data adaptively.

Table 5. The feature representation weights learnt by HB-MKL

	KTH	UCF	HMDB51
Trajectory	0.19	0.23	0.20
HOF	0.23	0.22	0.21
HOG	0.12	0.25	0.24
MBHx	0.23	0.16	0.17
MBHy	0.23	0.14	0.18

5 Conclusion

In this paper, we have presented an efficient feature fusion framework based on hierarchical Bayesian multiple kernel learning for action recognition. The method is able to integrate different features in an informative way by evaluating the relative importance of every feature and finally learns the optimum kernel combination of the multiple feature kernels. We have carried out a set of experiments on three human action datasets to evaluate the effectiveness of our approach, and the results have demonstrated that the proposed approach generally outperforms the state-of-the-art methods in terms of classification accuracy for action recognition.

Acknowledgments. This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the Natural Science Foundation of China (Grant No. 61472421, U1636218, 61472420, 61370185, 61170193, 61472063) , the Strategic Priority Research Program of the CAS (Grant No. XDB02070003), the Natural Science Foundation of Guangdong Province (Grant No. S2013010013432, S2013010015940), and the CAS External cooperation key project.

References

1. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008) 1–8
2. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: ECCV. (2006) 428–441
3. Scovanner, P., Ali, S., Shah, M.: A 3-dimensional sift descriptor and its application to action recognition. In: ACM MM. (2007) 357–360
4. Klaser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3d-gradients. In: BMVC. (2008)
5. Willems, G., Tuytelaars, T., Gool, L.V.: An efficient dense and scale-invariant spatio-temporal interest point detector. In: ECCV. (2008) 650–663
6. Sun, J., Wu, X., Yan, S., Cheong, L.F., Chua, T.S., Li, J.: Hierarchical spatio-temporal context modeling for action recognition. In: CVPR. (2009) 2004–2011
7. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: CVPR. (2011) 3169–3176

8. Tian, Y., Cao, L., Liu, Z., Zhang, Z.: Hierarchical filtered motion for action recognition in crowded videos. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on **42** (2012) 313–323
9. Ullah, M.M., Parizi, S.N., Laptev, I.: Improving bag-of-features action recognition with non-local cues. In: BMVC. (2010) 95.1–95.11
10. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. IJCV **73** (2007) 213–238
11. Girolami, M., Rogers, S.: Hierarchic bayesian models for kernel learning. In: ICML. (2005) 241–248
12. Damoulas, T., Girolami, M.A.: Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection. Bioinformatics **24** (2008) 1264–1270
13. Gönen, M.: Bayesian efficient multiple kernel learning. In: ICML. (2012) 1–8
14. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: ICCV. (2013) 3551–3558
15. Perronnin, F., Dance, C.: Fisher kernels on visual vocabularies for image categorization. In: CVPR. (2007) 1–8
16. Beal, M.J.: Variational algorithms for approximate Bayesian inference. University of London London (2003)
17. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local svm approach. In: ICPR. Volume 3. (2004) 32–36
18. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR. (2008) 1–8
19. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: Hmdb: A large video database for human motion recognition. In: ICCV. (2011) 2556–2563
20. Sheng, B., Yang, W., Sun, C.: Action recognition using direction-dependent feature pairs and non-negative low rank sparse model. Neurocomputing **158** (2015) 73–80
21. Zhang, H., Zhou, W., Reardon, C., Parker, L.E.: Simplex-based 3d spatio-temporal feature description for action recognition. In: CVPR. (2014) 2067–2074
22. Wu, J., Zhang, Y., Lin, W.: Towards good practices for action video encoding. In: CVPR. (2014) 2577–2584
23. Sun, L., Jia, K., Chan, T., Fang, Y., Wang, G., Yan, S.: Dl-sfa: Deeply-learned slow feature analysis for action recognition. In: CVPR. (2014) 2625–2632
24. Yang, X., Tian, Y.L.: Action recognition using super sparse coding vector with spatio-temporal awareness. In: ECCV. (2014) 727 – 741
25. Veeriah, V., Zhuang, N., Qi, G.: Differential recurrent neural networks for action recognition. In: ICCV. (2015) 4041–4049
26. Lan, T., Zhu, Y., Zamir, A.R., Savarese, S.: Action recognition by hierarchical mid-level action elements. In: ICCV. (2015) 4552 – 4560
27. Shao, L., Liu, L., Yu, M.: Kernelized multiview projection for robust action recognition. IJCV (2015) 1–15
28. Wang, D., Shao, Q., Li, X.: A new unsupervised model of action recognition. In: ICIP. (2015) 1160–1164
29. Liu, A.A., Su, Y.T., Nie, W.Z., Kankanhalli, M.: Hierarchical clustering multi-task learning for joint human action grouping and recognition. T-PAMI (2016) 1–14
30. Liu, L., Shao, L., Li, X., Lu, K.: Learning spatio-temporal representations for action recognition: A genetic programming approach. IEEE Transactions on Cybernetics **46** (2016) 158–170