Localized Multiple Kernel Learning for Realistic Human Action Recognition in Videos

Yan Song, Yan-Tao Zheng, Sheng Tang, Xiangdong Zhou, Yongdong Zhang, Shouxun Lin, and Tat-Seng Chua

***Abstract*—Realistic human action recognition in videos has been a useful yet challenging task. Video shots of same actions may present huge intra-class variations in terms of visual appearance, kinetic patterns, video shooting, and editing styles. Heterogeneous feature representations of videos pose another challenge on how to effectively handle the redundancy, complementariness and disagreement in these features. This paper proposes a localized multiple kernel learning (L-MKL) algorithm to tackle the issues above. L-MKL integrates the localized classifier ensemble learning and multiple kernel learning in a unified framework to leverage the strengths of both. The basis of L-MKL is to build multiple kernel classifiers on diverse features at subspace localities of heterogeneous representations. L-MKL integrates the discriminability of complementary features locally and enables localized MKL classifiers to deliver better performance in its own region of expertise. Specifically, L-MKL develops a locality gating model to partition the input space of heterogeneous representations to a set of localities of simpler data structure. Each locality then learns its localized optimal combination of Mercer kernels of heterogeneous features. Finally, the gating model coordinates the localized multiple kernel classifiers globally to perform action recognition. Experiments on two datasets show that the proposed approach delivers promising performance.**

***Index Terms*—Action recognition, localized classifier, multiple kernel learning.**

I. Introduction

RECOGNITION of human actions, like kissing, fighting, and so on, in videos has become an increasingly popular research topic, due to its wide applications in many vision tasks, such as event detection in surveillance, sports videos and movies, and so on. Though previous research efforts [1]–[3] have achieved promising performance in constrained videos, recognition accuracy remains unsatisfactory in unconstrained videos, such as movies and web videos. This is so because constrained videos tend to have clean background with predictable human actions, while unconstrained videos are real-life videos with large variations in visual contents.

This paper focuses on realistic human action recognition in unconstrained movie and web videos. The difficulties of realistic action recognition lie mainly in two facts. First, unconstrained videos possess huge variations in visual appearance, kinetic patterns, camera shooting and video editing styles, and so on. As shown in Fig. 1, this diversity renders video shots of the same human action to have varying motion and visual patterns. Consequently, the resulting huge intra-class variation hinders the performance of most statistical learning approaches. Second, different from images, video is a dynamic media carrier that possesses heterogeneous information from multiple channels, such as visual, audio, motion, and so on. The information from these channels can be regarded as different representations or views of the same action semantic. How to effectively handle the redundancy, complementariness and disagreement in heterogeneous feature representations poses another challenge to the recognition system.

To address the issues above, we propose a localized multiple kernel learning (L-MKL) method to integrate the localized classifier ensemble learning and multiple kernel learning in a unified framework. As shown in Fig. 2, the main idea is first to transform the global non-linear multi-representation input space into a set of localities with simpler data structure, and then build localized multiple kernel classifiers at subspaces of heterogeneous representations. The proposed L-MKL brings in two advantages. First, in the subspace of simpler complexity, the localized classifier is expected to deliver better accuracy in its own region of expertise; and the aggregation of these classifiers shall deliver superior performance to better tackle the issue of huge visual and motion variations in videos. Recent studies have demonstrated, both theoretically [4] and empirically [5], [6], that a good ensemble of localized classifiers can outperform a single classifier learned over the entire dataset. Second, the multiple kernel learning (MKL) [34]–[37] enables our L-MKL method to integrate the discriminability of complementary features locally, so as to better leverage heterogeneous feature representations of videos. Moreover, the formulation of a semi-definite programming task or sequential

Manuscript received August 17, 2010; revised November 24, 2010; accepted December 26, 2010. Date of publication March 17, 2011; date of current version September 2, 2011. This work was supported by the National Basic Research Program of China (973 Program, 2007CB311105), by the National Nature Science Foundation of China, under Grant 60873165, and by the Co-Building Program of Beijing Municipal Education Commission. This paper was recommended by Associate Editor C.-W. Lin.

Y. Song is with the Laboratory of Advanced Computing Research, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: songyan@ict.ac.cn).

Y.-T. Zheng is with the Institute for Infocomm Research, A*STAR, Singapore (e-mail: yzheng@i2r.a-star.edu.sg).

S. Tang, Y. Zhang, and S. Lin are with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China (e-mail: ts@ict.ac.cn; zhyd@ict.ac.cn; sxlin@ict.ac.cn).

X. Zhou is with the School of Computer Science and Technology, Fudan University, Shanghai 200433, China (e-mail: xdzhou@fudan.edu.cn).

T.-S. Chua is with the School of Computing, National University of Singapore, Singapore (e-mail: chuats@comp.nus.edu.sg).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2011.2130230

TABLE I
DESCRIPTION OF SYMBOLS IN PROBLEM FORMULATION

Symbol	Description	Parameter Estimating Step
K	The number of localized classifiers	–
M	The number of kernels	–
$\Psi_k\ (x)$	The gating function of x to the kth locality	Learning locality gating model
π_k	The kth locality of input space	Learning locality gating model
β_m^k	The mth kernel weight of the kth localized classifier	Learning multiple kernel classifier
ω_m^k	The weight coefficient of the mth kernel of the kth localized classifier	–
b^k	The coefficient of the kth localized classifier	Learning multiple kernel classifier
$\Phi_m(x)$	The mapping function related to the mth kernel	–
α^k	The Lagrange multiplier of the kth localized classifier	learning multiple kernel classifier

Fig. 1. Examples of intra-class diversity of human action in movies.

minimal optimization and so on [9], [10] enables the multiple kernel learning to converge fast.

In particular, the proposed L-MKL method first borrows the idea of multi-view clustering [18] to develop a locality gating model, which partitions the input space of heterogeneous representations to a set of localities of simpler data structure. The locality gating model exploits the expectation maximization (EM) algorithm to achieve maximal agreement between independent hypotheses of different representations. Then, based on the locality definition from previous step, the proposed approach learns localized classifiers with optimal combination of Mercer kernels of heterogeneous feature representations. Finally, it performs recognition by a global coordination of localized MKL classifiers.

The main contribution of this paper is that we propose a localized multiple kernel learning scheme for human action recognition in unconstrained real-life videos. The proposed approach integrates the localized classifier ensemble learning and multiple kernel learning in a unified framework. Testing on Hollywood-2 [7] and YouTube [8] datasets shows that the proposed approach achieves promising results and outperforms existing approaches with considerable margin.

TABLE II
ALGORITHM 1

Learning locality gating model via mixture of Gaussians EM

Input: $\{x_i^{(v)}, y_i \in \{\pm 1\} | i = 1, 2, \ldots N; v = 1, 2, \ldots V\}$, K
Output: $\{\pi_k\}_{k=1,\ldots,K}$, $\{\Psi_k\}_{k=1,\ldots,K}$

1.Use positive training set $\{x_i | y_i = +1, i = 1, 2, \ldots, N\}$ to do multi-view EM:

a. Randomly select K samples to be the centers and run k-means algorithm to obtain the initial $\gamma^{(1,0)}$ for $v = 1, n = 0$

b. Do the following loop until the stopping criterion is met:

For $v = 2, \ldots, V, 1, 2, \ldots$:

i) $n = n + 1$

ii) M-step: Compute model parameter $\Theta^{(v,n)}$ by (4), (5), (6)

iii) E-step: Update the hidden variable $\gamma^{(v,n)}$ by (7)

end

2. Obtain the localities of the input space $\{\pi_k\}_{k=1,\ldots,K}$ by (11)

3. Obtain the gating functions $\{\Psi_k(x_t)\}_{k=1,\ldots,K}$ by (12)

The rest of this paper is organized as follows. We first review the related work in Section II and elaborate on the details of the proposed L-MKL method in Section III. Analysis on L-MKL is presented in Section IV. Experiments are described in Section V. Finally, we present conclusive remarks along with discussion for future work in Section VI.

II. RELATED WORK

Earlier methods on human action recognition in videos mainly adopted holistic features like silhouette or shape [27], and human body model [28]. Unfortunately, holistic features based methods depend highly on the performance of segmentation and tracking, which may not deliver satisfactory results in realistic videos due to occlusion and cluttered background. These years, researchers tend to see videos as volumes instead of sequences of frames which induce methods from volumetric view [42], [43]. Particularly, the success of local features, like SIFT [22], in object recognition has thrusted the development of action recognition based on local spatial temporal (ST) features for action recognition [3], [13], [29]–[31]. To encode structure information, Kovashka and Grauman [32] proposed a hierarchy of discriminative space-time neighborhood features. Bregonzio *et al.* [33] used clouds of ST interest points to recognize actions which exploited the global ST distribution of interest points. Liu *et al.* [7], [12] combined static and motion features to recognize realistic action in web videos. Niebles *et al.* [1] proposed an unsupervised learning method by

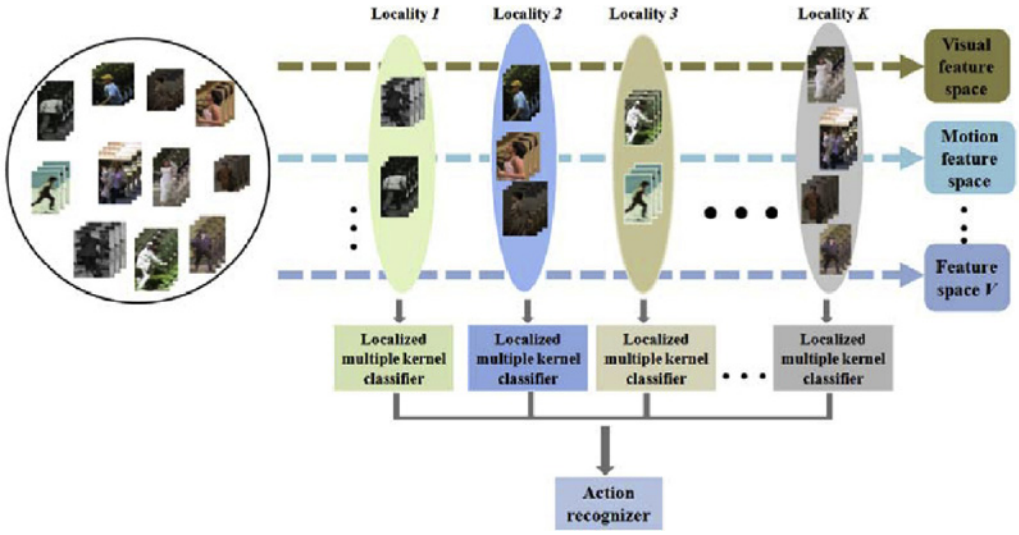

Fig. 2. Proposed L-MKL is to build localized classifiers on multiple features at local subspaces of heterogeneous representations.

TABLE III
ALGORITHM 2

Localized multiple kernel learning

Input: $\{x_i^{(v)}, y_i \in \{\pm 1\} | i = 1, 2, \ldots N; v = 1, 2, \ldots V\}$, K
Output: $F(x_t)$

1. Run algorithm 1 to obtain the localities $\{\pi_k\}_{k=1,\ldots,K}$ and the gating functions $\{\Psi_k\}_{k=1,\ldots,K}$
2. **For** $k = 1, 2, \ldots K$, learn decision function $f_k(x)$ on locality π_k:
   a. Set $\beta_m = 1/M$ for $m = 1, \ldots, M$
   b. **Repeat:**
      Compute $J(\beta)$ by an SVM solver
      Compute $\partial J/\partial \beta_m$, the largest component β^* and descent direction D
      While $J(\beta^*) < J(\beta)$
         Do{descent direction update}
         Update D^*, β^*
         Compute $J(\beta^*)$ by an SVM solver
      End While
      Until stopping criterion is met
   c. Line search along D to update β.
   End for
3. Obtain the discriminant function $F(x_t)$ by (3).

probabilistic latent semantic analysis model and latent Dirichlet allocation model for human action recognition. Hu *et al.* [14] focused on classification model by devising a multiple instance learning method for action detection in surveillance videos. For action recognition in real movie videos, Ivan *et al.* [11] explored text information while Marcin *et al.* [8] utilized scene classification to help recognize actions. Gilbert *et al.* [44] proposed a hierarchical compound feature for action recognition. Different from the aforementioned methods, this paper focuses on the issues of huge intra-class variation of human actions and multiple feature representations of videos. It attempts to tackle these two issues simultaneously by integrating ensemble learning and multiple kernel learning in a unified framework.

MKL [10], [16], [20], [23] has been recently used in many applications to fuse features from multiple modalities, such as genomic data fusion [15], object classification [34], [35], [37], and object detection [36]. Varma and Ray [34] combined heterogeneous local descriptors to seek optimal tradeoff between representation discrimination and invariance. Nakajima *et al.* [35] adopted a recently developed non-sparse MKL for combining information from various image descriptors for object classification task. Kembhavi *et al.* [37] proposed an incremental multiple kernel learning method for object recognition. In this paper, MKL is used to fuse heterogeneous features. However, in contrast to the existing MKL approaches, the proposed L-MKL fuses heterogeneous features at local subspaces of various video representations, in the spirit of ensemble learning.

An ensemble is a scheme to combine many weak learners to produce a strong learner. Ensemble learning methods build a set of classifiers and predict the test data by coordinating the results of the ensemble of classifiers for better performance. Empirically, ensemble methods tend to yield better results when there is a significant diversity among the constituent classifiers [45]. Existing methods of constructing ensembles include Bayesian voting, manipulation on the training example and input features, and so on [39]. Jacobs *et al.* [38] proposed an adaptive mixture of local experts which are separate networks. Bagging and boosting schemes [17], [24], [25] combined weak classifiers to generate a strong one. In our framework, L-MKL adopts a set of localized classifiers trained on localities of input space to perform ensemble learning. Tang *et al.* proposed a latent Dirichlet allocation-support vector machine (LDA-SVM) by using LDA to cluster samples into topics and train the data in each topic [41]. In the context of kernel methods, localized classifier learning was performed by assigning different weights to kernels in different regions of the input space [16], [40]. However, different from ensemble learning methods, these localized-classifier methods only learn a single classifier in nature by adapting the kernel combinations with data space locality.

III. ALGORITHM

A. Preliminaries and Problem Formulation

In the task of human action recognition, a video shot is processed to detect the existence of a set of human motion or gestures, like "jumping," "kissing," and so on. To

represent a video shot, several feature representations are available, such as static visual features extracted from shot key frames [22] and dynamic local ST features extracted from video sequence [3]. Let $\{x^{(v)} jx^{(v)} \in R^{Dv},\ v = 1, 2, \ldots, V\}$ represent V heterogeneous feature representations (or views) for a video shot x, where D_v represents the dimensionality of feature v. For each action class, we have a training set $X = \{x_i^{(v)} | x_i^{(v)} \in R^{Dv}, v = 1, 2, \ldots, V, i = 1, 2, \ldots, N\}$ with label $Y = \{y_i \in \{\pm 1\},\ i = 1, 2, \ldots N\}$, where N is the number of training samples. Human action recognition can then be naturally formulated as a classification task.

Here, we adopt the binary classification formulation in the framework of support vector machine (SVM). We aim to learn a classifier with a discriminant function $F(x_t)$ for a test sample x_t

$$F(x_t) = \sum_{k=1}^{K} \psi_k(x_t) \left(\sum_{m=1}^{M} \beta_m^k \left\langle \omega_m^k, \Phi_m(x_t) \right\rangle + b^k \right)$$
$$\beta_m^k \geq 0 \quad \sum_{m=1}^{M} \beta_m^k = 1 \quad \forall k. \tag{1}$$

The discriminant function (or action classifier) $F(x_t)$ is an ensemble of K localized classifiers that are built in multi-representation local subspaces of V heterogeneous features.

By solving the primal SVM problem, we can obtain ω_m as follows:

$$\omega_m = \sum_{i=1}^{N} \alpha_i y_i \Phi_{\mathrm{m}}(x_i). \tag{2}$$

By plugging in ω_m, (1) can be rewritten as

$$F(x_t) = \sum_{k=1}^{K} \psi_k(x_t) \left(\sum_{m=1}^{M} \beta_m^k \sum_{i \in \pi_k} \alpha_i^k y_i \left\langle \Phi_m(x_t), \Phi_m(x_i) \right\rangle + b^k \right)$$
$$\beta_m^k \geq 0 \quad \sum_{m=1}^{M} \beta_m^k = 1 \quad \forall k. \tag{3}$$

Learning action classifier, namely (3), can be decomposed into two steps: 1) estimating locality gating model, and 2) computing multiple kernel classifier parameters in multi-representation data space localities. In the first step, estimating the locality gating model includes obtaining the localities $\{\pi_k\}_{k=1,\ldots,K}$ of multi-representation data space and the gating functions $\{\Psi_k\}_{k=1,\ldots,K}$. The localities $\{\pi_k\}_{k=1,\ldots,K}$ determine the subspaces of the input space and the gating functions $\{\Psi_k(x_t)\}_{k=1,\ldots,K}$ determine the weights of localized classifiers for a test sample x_t. In the second step, multiple kernel classifier parameters $\{\{\beta_m^k\}, \{a_i^k\}, \{b^k\}\}$ are inferred to characterize the local classifier at each locality. For reference ease, the variables in the L-MKL model are listed in Table I, together with their descriptions.

B. Learning Locality Gating Model

The main task of learning locality gating model is to partition the positive training set of multiple representations by maximizing the agreement between independent hypotheses of different representations. The local consensus between multiple feature representations enables the resulting localities to be of simpler complexities, and thus, facilitates better localized classifier learning. In the spirit of multi-view clustering [18], the model adopts the expectation-maximization (EM) algorithm to achieve the consensus of different hypotheses, and furthermore, to infer the gating functions for test samples. The premise here is that the disagreement between two independent hypotheses is an upper bound on the error risk of either hypothesis [19]. The spirit of multi-view EM algorithm is that different representations exchange the expected values for hidden variables in each iteration step of EM process, on which they find the locality model parameters that maximize the likelihood.

Here, we assume that a sample x is generated by a mixture of K Gaussian distributions [21] with parameters $\Theta = \{\rho_k, \mu_k, \Sigma_k\}_{k=1,\ldots,K}$ in each representation space, where ρ is the mixing weights, μ is the mean, and Σ is the variance. Let $\gamma(i, k)$ denote the probability that the ith sample x_i is generated by the kth Gaussian. We can therefore obtain our multi-representation mixture of Gaussian EM algorithm. In the nth M-step with representation v, the parameters Θ of Gaussian models are estimated by the expectation value of hidden variable $\gamma^{(v-1,n-1)}$ obtained in the $(n-1)$th E-step computed in representation $v-1$, as defined by (4)–(6)

$$\mu_k^{(v,n)} = \frac{1}{\sum_{i=1}^{N} \gamma^{(v-1,n-1)}(i,k)} \sum_{i=1}^{N} \gamma^{(v-1,n-1)}(i,k) x_i^{(v)} \tag{4}$$

$$\Sigma_k^{(v,n)} = \frac{\sum_{i=1}^{N} \gamma^{(v-1,n-1)}(i,k)(x_i^{(v)} - \mu_k^{(v,n)})(x_i^{(v)} - \mu_k^{(v,n)})^T}{\sum_{i=1}^{N} \gamma^{(v-1,n-1)}(i,k)} \tag{5}$$

$$\rho_k^{(v,n)} = \frac{1}{N} \sum_{i=1}^{N} \gamma^{(v-1,n-1)}(i,k). \tag{6}$$

In the nth E-step, the expectation value of hidden variable γ is updated by the parameters computed in the nth M-step in representation v as follows:

$$\gamma^{(v,n)}(i,k) = \frac{\rho_k^{(v,n)} p(x_i^{(v)}; \mu_k^{(v,n)}, \Sigma_k^{(v,n)})}{\sum_{k'=1}^{K} \rho_{k'}^{(v,n)} p(x_i^{(v)}; \mu_{k'}^{(v,n)}, \Sigma_{k'}^{(v,n)})} \tag{7}$$

$$p(x; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2} \det(\Sigma)^{1/2}} \exp\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\}. \tag{8}$$

Then M step in view $v+1$ is executed with $\gamma^{(v,n)}$. In this way, expected values for hidden variables are interchanged among different views.

The iteration is terminated when the convergence criterion is met. Convergence of the algorithm can be determined by observing the change of log-likelihood of data in each representation

$$\log P(X^{(v)} | \Theta^{(v)}) = \sum_{i=1}^{N} \log \left(\sum_{k=1}^{K} \gamma(i,k) p(x_i^{(v)}; \mu_k^{(v)}, \Sigma_k^{(v)}) \right). \tag{9}$$

When the iteration is terminated, we adopt the final value of hidden variable γ to define the partition $\{\pi_k^+\}_{k=1,\ldots,K}$ of the positive training set as

$$\pi_k^+ = \{x_i : k = \arg\max_{k'} (\sum_{v=1}^{V} \gamma^{(v)}(i,k'))\}. \tag{10}$$

TABLE IV
OPTIMAL VALUE OF K FOR ACTION CATEGORIES IN HOLLYWOOD-2 DATASET

ActionCategory	Answer Phone	Drive Car	Fight	Hand Shake	Hug	Kiss	SitDown	Stand Up	Eat	Run	GetOutCar	SitUp
Optimal K	2	5	4	3	4	4	3	5	5	4	3	5

Then the kth locality π_k of the input space is defined as

$$\pi_k = \pi_k^+ \bigcup \{x_i | y_i = -1, i = 1, 2, ...N\}. \tag{11}$$

We determine the gating function $\Psi_k(x_t)$ for a test sample x_t in (1) and (2) by the final GMM parameters as

$$\Psi_k(x_t) = \frac{p(k|x_t)}{\sum_{k'=1}^{K} p(k'|x_t)}$$
$$p(k|x_t) = \sum_{v=1}^{V} \frac{\rho_k^{(v)} p(x_t^{(v)}; \mu_k^{(v)}, \Sigma_k^{(v)})}{\sum_{k'=1}^{K} \rho_{k'}^{(v)} p(x_t^{(v)}; \mu_{k'}^{(v)}, \Sigma_{k'}^{(v)})}. \tag{12}$$

Intuitively, the gating function $\Psi_k(x_t)$ determines the membership of sample x_t in each locality with consensus of multiple feature representations.

The algorithm for learning the locality gating model via the mixture of Gaussian EM is summarized in Algorithm 1.

C. Learning Multiple Kernel Classifier

After learning the locality gating model that defines the locality of the input space, we learn the multiple kernel classifier parameters $\{\{\beta_m^k\}, \{\alpha_i^k\}, \{b^k\}\}$ for localized classifier in each locality. Specifically, the learning procedure is to obtain a decision function in (13) for each locality (the locality index k is omitted in the following)

$$f(x) = \sum_i \alpha_i y_i K(x, x_i) + b \tag{13}$$

$$K(x, x_i) = \sum_{m=1}^{M} \beta_m K_m(x, x_i) \quad \beta_m \geq 0 \quad \sum_{m=1}^{M} \beta_m = 1 \tag{14}$$

where K_m is a positive definite kernel. A weighted 2-norm regularization [20] is explored to solve the problem above. The optimization problem then becomes

$$\min(J(\beta)) = \min_{\{f\},b,\xi} \frac{1}{2} \sum_m \frac{1}{\beta_m} ||f_m||_{H_m}{}^2 + C \sum_i \xi_i$$
$$\text{s.t.} \quad y_i \sum_m f_m(x_i) + y_i b \geq 1 - \xi_i$$
$$\xi_i \geq 0 \quad \forall i \quad \beta_m \geq 0 \quad \sum_{m=1}^{M} \beta_m = 1 \tag{15}$$

where f_m belongs to a reproducing kernel Hilbert space associated with a kernel K_m. Assuming that each gram matrix $K_m(x_i, x_j)$ is positive definite, the optimization function J is convex and differentiable. The approach used to solve this optimization problem is a reduced gradient method. The reduced gradient of $J(\beta)$ is

$$\nabla J_t = \frac{\partial J}{\partial \beta_t} - \frac{\partial J}{\partial \beta_s}, t \neq s$$
$$\nabla J_s = \sum_{t \neq s} \left(\frac{\partial J}{\partial \beta_s} - \frac{\partial J}{\partial \beta_t} \right) \tag{16}$$

where s is chosen as the index of the largest component of β. Algorithm 2 summarizes the overall flow of the proposed L-MKL algorithm, with emphasis on the part of multiple kernel classifier learning.

With a set of local multiple-kernel classifiers, the prediction on a testing sample is decided by the discriminant function $F(x_t)$ in (1), which is effectively a weighted average of decision scores of individual local classifiers. The weights are determined by the gating functions of each locality, as introduced in Part A, Section III. Fig. 3 illustrates the training and testing processes of our algorithm.

IV. ANALYSIS OF L-MKL

A. Parameterized Coordination of Local Models

The proposed L-MKL method can be regarded as a global coordination of locally linear discriminant models. It regulates the coordination with a global parameterization on the agreement of hypothesis of different data representations. This parameterization is solved by the expectation-maximization algorithm in a multi-view clustering approach. The coordination regulates the local multiple-kernel classifiers in two aspects: 1) it defines how the data samples are grouped into localities to train a localized MKL classifier, and 2) it governs how the predictions of localized MKL classifiers on a testing sample are combined. The global parameterization on multiple data representation furnishes the proposed L-MKL with two desirable properties that distinguish it from most traditional localized model learning and multiple kernel learning schemes. First, the partitioning of input space takes into account the heterogeneity and disagreement of multiple representations, which builds the basis for localized multiple kernel learning at the later stage. Second, the approach allows multiple kernel learning to be performed in a distributed fashion. This property is favorable, especially when the efficiency in multiple kernel learning is critical on large-scale dataset.

B. Computational Complexity

The computational complexity of L-MKL lies in two parts: locality gating model learning and multiple kernel parameter learning. For locality gating model learning, the complexity of EM algorithm is $O(K \times N_k)$, where K is the number of localities, and N_k is the number of data samples at the

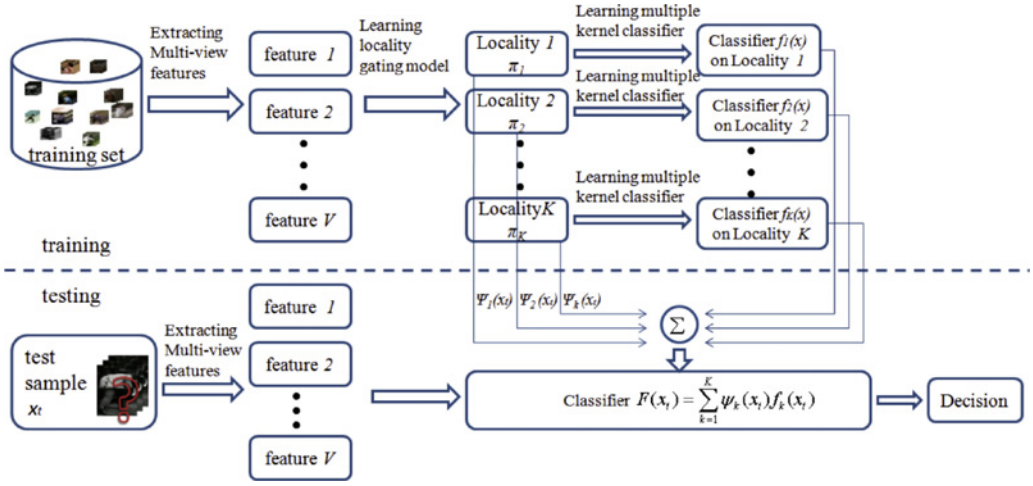

Fig. 3. Flowchart of the L-MKL algorithm.

current locality k and it can be approximated to be N/K. To learn the localized MKL classifier, each locality needs to solve the 2-norm regularization problem in SVM with combined kernel and calculate the gradient $J(\beta)$ for a few iterations. Compared to the complexity of solving a 2-norm regularization, the calculation of gradient $J(\beta)$ is trivial. At each iteration, determining gradient descent update requires additional optimization. The number of iterations l depends on the training data and gradient parameter. Therefore, the complexity of multiple kernel parameter learning at each locality depends mainly on SVM optimization $O(N_k^2)$. Consequently, the complexity of the whole multiple kernel parameter learning process is $l \times K \times O(N_k^2) = l \times O(N^2/K)$. As $K << N$, the computational complexity of the L-MKL algorithm is then $l \times O(N^2/K)$.

C. Conditional Independence of Feature Representations

One issue remains open in the proposed algorithm. The locality gating model learning implicitly assumes that different feature representations are conditionally independent, given their class label, namely $p(x_i^1, \ldots, x_i^v, \ldots, x_i^V | y_i) = \Pi p(x_i^v | y_i)$. The problem now is how practical this independence assumption is. Given the fact that it is, in general, infeasible to infer the joint density of different representations $p(x_i^1, \ldots, x_i^v, \ldots, x_i^V | y_i)$ from marginal densities, our concern becomes how this conditional independence assumption affects the final recognition performance. Our experimental result shows that the proposed algorithm delivers promising performance and outperforms other existing approaches with considerable margins. We therefore conjecture that the conditional independence assumption simplifies the model and permits better generalization.

V. Experiments

A. Datasets and Experimental Setup

1) *Datasets and Evaluation Criterion:* We test our algorithm on two datasets: Hollywood-2 [8] and YouTube dataset [7]. The Hollywood-2 dataset consists of 12 action categories, such as "answer phone," "drive car," and so on, and 2571 real-life movie videos in total. Fig. 4(a) shows some key frames of videos in the dataset. For benchmark purpose, we follow the setup of [8] to utilize 1707 video clips (823 for training and 884 for testing). We adopt average precision (AP) for each class and average AP (AAP) on all classes as the evaluation criterion. The YouTube dataset contains 1168 video clips of 11 categories, such as basketball shooting, volleyball spiking, trampoline jumping, and so on, as shown in Fig. 4(b). Each category is divided into 25 relatively independent groups. YouTube dataset is reported to be one of the most extensive realistic action dataset of web videos [7]. For YouTube dataset, we adopt recognition accuracy as evaluation criterion as [7] and extend our algorithm for multi-class classification in a one-versus-all setting.

Fig. 4. (a) Examples of the key frames in the Hollywood-2 datasets. (b) Examples of the key frames in the YouTube datasets.

2) *Features and Kernels:* Two types of features are employed. The first feature is a static feature that encodes the appearance information. The second feature is a dynamic

feature that encodes the motion information of the actions. For static appearance features, we exploit the bag-of-words representation, based on scale invariant feature transform (SIFT) descriptors [22], as it has been reported to deliver good practical performance. SIFT regions are extracted by difference of Gaussian detector. The SIFT feature in a video is represented by a histogram.

For the dynamic feature, we exploit the bag-of-words representation of local ST feature [3], which is a recently proposed feature for action recognition. Local ST features represent videos in a compact but discriminative manner by describing local cuboids at the most informative ST locations. The response function is $R = [I^* g_\sigma^* h_{ev}]^2 + [I^* g_\sigma^* h_{od}]^2$, where g_σ is Gaussian filter applied on spatial dimensions, h_{ev} and h_{od} are a quadrature pair of Gabor filters applied on temporal dimension. The parameters σ and τ are the spatial and temporal scales of the detector. Local 3-D cuboids of interest points are characterized by brightness gradient. PCA is used to reduce the dimension of feature of cuboids. The ST feature in a video is also represented by a histogram.

The distance measure for both features is χ^2 distance which is computed as

$$\mathrm{D}(h_i, h_j) = \frac{1}{2} \sum_{bin=1}^{H} \frac{[h_i(bin) - h_j(bin)]^2}{h_i(bin) + h_j(bin)} \tag{17}$$

where h_i and h_j are histograms of sample i and j, and H is the number of bins in the histogram.

For localized multiple kernel learning, we adopt polynomial (homogeneous), polynomial (inhomogeneous), Gaussian radial basis function (RBF) as kernels for each feature.

3) *Parameter Setting:* The spatial and temporal scales (σ and τ) are both set to 1.5 in our experiment. The dimension of cuboids feature is reduced to 100 by PCA. We apply k-means clustering to generate the codebooks of both features with cardinality equal to 500. The degree parameter d in polynomial kernel is set to 2. The parameter r in Gaussian RBF kernel $Kernel(x_i, x_j) = exp[-(1/r)^* D(x_i, x_j)]$ is set to the mean distance of all training samples as [11]. The regularization parameter C for MKL is set to 1000. For the number of localities K, we adopt the leave-one-out cross validation scheme.

B. The Hollywood-2 Dataset

1) *Locality Gating Model Learning:* Here, we test the effectiveness of the proposed locality gating model learning for multiple-representations input space partitioning. We perform locality gating model learning via Gaussian mixture EM to partition the input space into K localities. For efficiency purpose, we adopt PCA to reduce the feature dimensionality to 50 in the EM process. For each category, we set K from 2 to 6 and record K for the best performance in the cross validation. Table IV illustrates the value of K with optimal recognition performance. The value of optimal K ranging from 2 to 5 reveals the intra-class diversity, from the perspective of classification.

To test effectiveness of the proposed locality gating model, we compare the recognition performance of the following runs:

TABLE V
AVERAGE APS OF 4 RUNS FOR HOLLYWOOD-2 DATASET

Run	Run-1	Run-2	Run-3	Run-4
Average AP	0.3923	0.4020	0.4125	0.4314

1) *run-1*: random partitioning of the input space;
2) *run-2*: EM-based partitioning based on a single static feature;
3) *run-3*: EM-based partitioning based on a single dynamic feature;
4) *run-4*: proposed locality gating model based on multiple-feature representation.

For run-1, we randomly partition the input space into K localities (K is set to the number chosen in the cross validation step) and conduct our localized MKL method. For run-2 and run-3, similar gating model learning is applied, but on single static or dynamic feature only. Run-4 incorporates both the static and dynamic features in L-MKL as proposed in Section III. The AAP of all the runs are shown in Table V. The proposed gating model achieves the highest AAP of 0.4314, which outperforms run-1 by 10%, run-2 by 7.3%, and run-3 by 4.6% relatively. This demonstrates that a proper partitioning of multi-representation data space plays an important role in the performance of localized classifiers. By maximizing the agreement between independent hypotheses of different representations, the proposed locality gating model learning method enables the localized classifiers to leverage the discrimination of complementary features locally.

2) *L-MKL Classifier:* Next, we verify the effectiveness of the proposed L-MKL classifier. Specifically, we compare the performance of proposed L-MKL with a global MKL classifier [20] and two classical ensemble learning methods, i.e., Adaboost [24] and Bootstrap aggregating (Bagging) [25].

The global MKL classifier is trained over the entire dataset. Fig. 5 elaborates the APs of each category by L-MKL and global MKL. Overall, L-MKL achieves an AAP of 0.4314, which is 11.73% relatively higher than that of the global MKL classifier. Specifically, category "Run" has the largest AP improvement by 16.6% relatively. We attribute this improvement to the conjecture that this category might possess better salient sub-group locality structures. As each locality has simpler data structure, the localized MKL classifier tends to deliver more superior results. However, we also observe that for the "AnswerPhone" and "SitUp" categories, the proposed L-MKL method shows little improvement over the global classifier. Our postulation is that the sub-group locality distribution is not obvious in these groups, which renders locality learning ineffective.

Moreover, we compare the performance of our method with Adaboost [24] and Bootstrap aggregating (Bagging) [25]. Adaboost and Bagging are widely used ensemble learning methods that combine a set of local classifiers for final classification. Taking C4.5 [17] as weak classifier, Adaboost and Bagging are trained on each of static and motion feature, respectively. Table VI summarizes the AAP of Adaboost and Bagging on single feature. Fig. 5 illustrates the detailed AP

TABLE VI
BENCHMARK OF EXISTING METHODS ON HOLLYWOOD-2 DATASET

Method	Proposed L-MKL	Global MKL	Han *et al.* [26]	Wang *et al.* [13]	Adaboost on Motion Feature	Bagging on Motion Feature	Adaboost on Static Feature	Bagging on Static Feature
AAP	0.4314	0.3861	0.4212	0.45	0.2557	0.2602	0.1771	0.1724

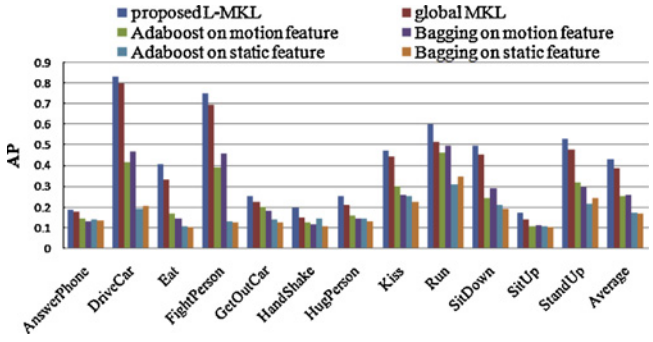

Fig. 5. Comparison of AP of the proposed L-MKL with the global MKL, Adaboost, and Bagging for all categories of Hollywood-2 dataset.

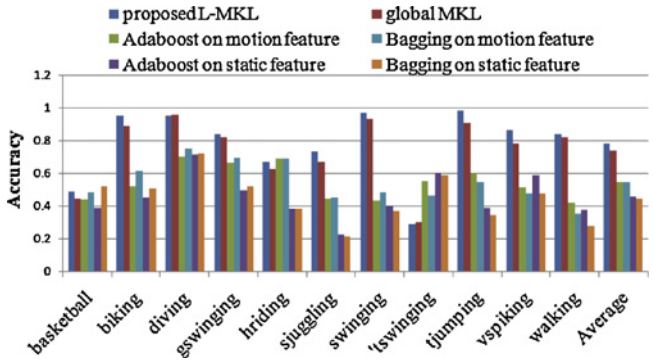

Fig. 6. Comparison of accuracies of the proposed L-MKL with the global MKL, Adaboost, and Bagging for all categories of the YouTube dataset.

TABLE VII
AVERAGE APS OF 4 RUNS FOR YOUTUBE DATASET

Run	Run-1	Run-2	Run-3	Run-4
Average accuracy	75.03%	75.86%	76.13%	77.91%

comparison on each category. As shown, the proposed L-MKL outperforms Adaboost and Bagging with considerable margins of 0.1712 and 0.259, respectively. We attribute this substantial improvement to the fact that the local classifiers in L-MKL are built on multiple features, while the classifiers in Adaboost and Bagging are on single feature only. In other words, the multiple kernel learning on heterogeneous features in L-MKL makes the ensemble learning more effective.

3) *Benchmark:* We also benchmark our method with the state-of-the-art methods [26] and [13] (we choose the result of the same dynamic feature as ours) in Table VI. As shown, our proposed method delivers better or comparable results.

C. The YouTube Dataset

1) *Locality Gating Model Learning:* Similar to the Hollywood-2 dataset, we first verify the effectiveness of the proposed locality learning on multiple representations, by performing the four runs listed in Section V-B. The average accuracies of the four runs are shown in Table VII, from which we obtain similar observation as in Hollywood-2 dataset. The proposed L-MKL method that uses all features achieves the best performance with an accuracy of 77.91%, which outperforms the other three runs.

2) *L-MKL Classifier:* Next, we compare the proposed localized MKL method with the global MKL scheme [20] and two ensemble learning methods, i.e., Adaboost [24] and Bagging [25] on YouTube dataset. The accuracies of the 11 categories are shown in Fig. 6. The average accuracy of the proposed method is 77.91%, which outperforms the global MKL by 5.37% relatively. Similarly, the proposed L-MKL method also outperforms Adaboost and Bagging by considerable margins.

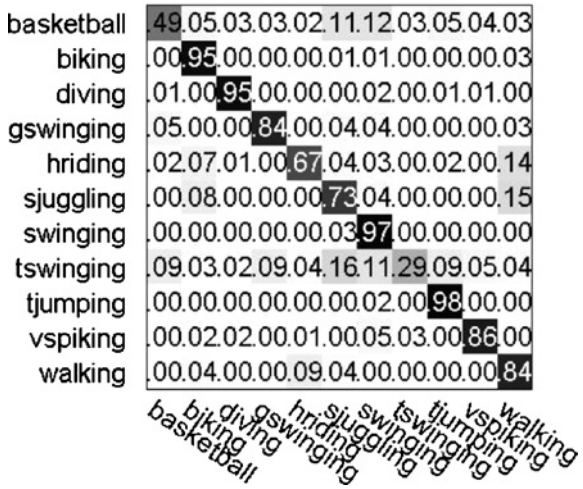

Fig. 7. Confusion matrix for the proposed L-MKL method on YouTube dataset.

To further analyze the difference between the global MKL and L-MKL classifier, we examine the accuracies of each category. As observed, the proposed L-MKL method outperforms the global MKL method on 9 out of 11 categories by a margin ranging from 2% to 7.96%. But two categories have the global MKL classifier outperforming the L-MKL, which are "diving" and "tennis swinging." After careful examination on videos of these two categories, we observe that the intra-class diversity of "diving" is mild, which renders the global classifier more effective. Moreover, the "tennis swinging" category is observed to share similar motion patterns with "swinging" and "sjuggling" categories. This low inter-class distance hinders the performance of localized classifiers, as data samples are more easily confused with ones of other classes at a locality. Fig. 7 shows the confusion matrix of the classification by the proposed L-MKL. The most confusing category pairs

TABLE VIII
BENCHMARK OF EXISTING METHODS ON YOUTUBE DATASET

Method	Proposed L-MKL	Global MKL	Liu *et al.* [7]	Adaboost on Motion Feature	Bagging on Motion Feature	Adaboost on Static Feature	Bagging on Static Feature
Average accuracy	77.91%	73.94%	71.21%	54.28%	54.36%	45.54%	44.49%

are "soccer_juggling (sjuggling)" and "walking," and "tennis swinging (tswinging)" and "soccer_juggling (sjuggling)."

3) *Benchmark:* Overall, the proposed L-MKL algorithm achieves an average accuracy of 77.91%, which is 6.7% higher than the average accuracy of [7] at 71.21%. The benchmark of YouTube dataset is shown in Table VIII.

VI. CONCLUSION

Recognizing human actions, like fighting, biking, and so on, in videos has been a popular research topic, due to its significance to many vision and multimedia applications. Two issues, however, encumber effective action recognition, which are: huge intra-class variations of human actions and multiple heterogeneous feature representations of videos. In this paper, we proposed a L-MKL algorithm to tackle these two issues. The proposed algorithm integrates the localized classifier ensemble learning and multiple kernel learning in a unified framework. In the algorithm, multiple kernel classifiers are built locally on heterogeneous features at multi-representation data subspaces. By adapting kernel combinations to data space locality, L-MKL integrates the discriminability of complementary features locally, and enables localized MKL classifiers to deliver better performance in its own region of expertise. Specifically, in the spirits of the multi-view clustering, the proposed method develops a locality gating model to partition the input space of multiple feature representations into a set of localities based on EM algorithm. Each locality of input space then learns a localized optimal combination of kernels of heterogeneous features. Finally, the locality gating model coordinates the localized MKL classifiers globally to perform action recognition. Testing on Hollywood-2 and YouTube datasets showed that the proposed algorithm delivers state-of-the-art results and outperforms existing approaches with considerable margin. Future work includes how to incorporate the inter-relation among action classes in the recognition frame work, such as the motion tempo relation of "running" and "walking" categories.

REFERENCES

[1] J. C. Niebles, H. C. Wang, and F. F. Li, "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vis.*, vol. 79, no. 3, pp. 299–318, Sep. 2008.

[2] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. IEEE Int. Conf. Pattern Recognit.*, Sep. 2004, pp. 32–36.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proc. 14th Int. Conf. Comput. Commun. Netw.*, 2005, pp. 65–72.

[4] A. Krogh and J. Vedelsby, "Neural network ensembles, cross validation, and active learning," *Adv. Neural Inform. Process. Syst.*, vol. 7, pp. 231–238, 1995.

[5] D. W. Opitz and J. W. Shavlik, "Generating accurate and diverse members of a neural network ensemble," *Adv. Neural Inform. Process. Syst.*, vol. 8, pp. 535–541, 1996.

[6] D. W. Opitz and J. W. Shavlik, "Actively searching for an effective neural network ensemble," *Connection Sci.*, vol. 8, nos. 3–4, pp. 337–354, Dec. 1996.

[7] J. G. Liu, J. B. Luo, and M. Shah, "Recognizing realistic actions from videos 'in the wild,"' in *Proc. IEEE Int. Conf. Comput. Vision Pattern Recognit.*, Jun. 2009, pp. 1996–2003.

[8] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2929–2936.

[9] G. R. G. Lanckriet, M. Deng, N. Cristianini, P. Bartlett, L. E. Chaoui, and M. I. Jordan, "Learning the kernel matrix with semi-definite programming," *J. Mach. Learning Res.*, vol. 5, pp. 27–72, Dec. 2004.

[10] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proc. 21th Int. Conf. Mach. Learning*, 2004, p. 6.

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[12] J. G. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[13] H. Wang, M. M. Ullah, A. Kläser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *Proc. Br. Mach. Vis. Conf.*, Sep. 2009, p. 127.

[14] Y. Hu, L. L. Cao, F. J. Lv, S. C. Yan, Y. H. Gong, and T. S. Huang, "Action detection in complex scenes with spatial and temporal ambiguities," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep.–Oct. 2009, pp. 128–135.

[15] G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, no. 16, pp. 2626–2635, Nov. 2004.

[16] M. Gönen and E. Alpaydin, "Localized multiple kernel learning," in *Proc. 25th Int. Conf. Mach. Learning*, 2008, pp. 352–359.

[17] J. R. Quinlan, "Bagging, boosting, and c4.5," in *Proc. 13th Natl. Conf. Artif. Intell.*, 1996, pp. 725–730.

[18] S. Bickel and T. Scheffer, "Multi-view clustering," in *Proc. 4th IEEE Int. Conf. Data Mining*, Nov. 2004, pp. 19–26.

[19] S. Dasgupta, M. Littman, and D. McAllester, "PAC generalization bounds for co-training," in *Advances in Neural Information Processing Systems*, vol. 14. Cambridge, MA: MIT Press, 2002, pp. 375–382.

[20] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *J. Mach. Learning Res.*, vol. 9, pp. 2491–2521, Nov. 2008.

[21] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford, U.K.: Oxford University Press, 1995.

[22] D. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comp. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[23] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *J. Mach. Learning Res.*, vol. 7, pp. 1531–1565, Jul. 2006.

[24] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Mach. Learning*, vol. 37, no. 3, pp. 297–336, 1999.

[25] L. Breiman, "Bagging predictors," *Mach. Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[26] D. Han, L. F. Bo, and C. Sminchisescu, "Selection and context for action recognition," in *Proc. IEEE Int. Conf. Comput. Vision*, Sep.–Oct. 2009, pp. 1933–1940.

[27] A. Veeraraghavan, A. R. Chowdhury, and R. Chellappa, "Matching shape sequences in video with application in human movement analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1896–1909, Dec. 2005.

[28] H. J. Li, J. H. Tang, S. Wu, Y. D. Zhang, and S. X. Lin, "Automatic detection and analysis of player action in moving background sports video sequences," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 3, pp. 351–364, Mar. 2010.

[29] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Proc. Eur. Conf. Comput. Vision*, 2008, pp. 650–663.

[30] A. Kläser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3-D gradients," in *Proc. Br. Mach. Vis. Conf.*, 2008, pp. 995–1004.
[31] I. Laptev and T. Lindeberg, "Space-time interest points," in *Proc. Int. Conf. Comput. Vision*, 2003, pp. 432–439.
[32] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminant space-time neighborhood features for human action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1–8.
[33] M. Bregon, S. G. Gong, and T. Xiang, "Recognizing action as clouds of space-time interest points," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, Jun. 2009, pp. 1948–1955.
[34] M. Varma and D. Ray, "Learning the discriminative power-invariance tradeoff," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2007, pp. 1–8.
[35] S. Nakajima, A. Binder, C. Muller, W. Wojcikiewicz, M. Kloft, U. Brefeld, K.-R. Muller, and M. Kawanabe, "Multiple kernel learning for object classification," in *Proc. 12th Workshop Information-Based Induction Sci.*, 2009.
[36] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. IEEE Int. Conf. Comput. Vision*, May 2009, pp. 606–613.
[37] A. Kembhavi, B. Siddiquie, R. Miezianko, S. McCloskey, and L. S. Davis, "Incremental multiple kernel learning for object recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2009, pp. 638–645.
[38] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton, "Adaptive mixtures of local experts," *Neural Comput.*, vol. 3, no. 1, pp. 79–87, 1991.
[39] T. G. Dietterich, "Ensemble methods in machine learning," in *Proc. Workshop Multiple Classifers Syst.*, 2000, pp. 1–15.
[40] J. J. Yang, Y. N. Li, Y. H. Tian, L. Y. Duan, and W. Gao, "Group-sensitive multiple kernel learning for object categorization," in *Proc. IEEE Int. Conf. Comput. Vision*, May 2009, pp. 436–443.
[41] S. Tang, J. T. Li, M. Li, C. Xie, Y. Z. Liu, K. Tao, and S. X. Xu, "TRECVID 2008 high-level feature extraction by MCG-ICT-CAS," in *Proc. TRECVID Workshop*, 2008 [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html
[42] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 2, pp. 288–303, Feb. 2010.
[43] H. J. Seo and P. Milanfar, "Action recognition from one example," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 867–882, May 2011.
[44] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 883–897, May 2011.
[45] L. I. Kuncheva and C. J. Whitaker, "Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy," *Mach. Learning*, vol. 51, no. 2, pp. 181–207, May 2003.

Yan Song received the B.S. degree from the Nanjing University of Science and Technology, Nanjing, Jiangsu, China, in 2005. She is currently pursuing the Ph.D. degree from the Multimedia Computing Group, Laboratory of Advanced Computing Research, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

Her current research interests include multimedia information processing, in particular, video content analysis and understanding. The main focus is on the human action recognition and event detection of video.

Yan-Tao Zheng received the B.Eng. (with first class honors) degree from Nanyang Technological University, Singapore, and the Ph.D. degree from the National University of Singapore, Singapore.

He is currently a Research Engineer with the Institute for Infocomm Research, A*STAR, Singapore. His current research interests include geo-mining in multimedia, image annotation, and video search.

Dr. Zheng is the recipient of a number of international awards, including Champion of Star Challenge, Microsoft Research Fellowship, IBM Waston Emerging Multimedia Leaders, and so on. During his attachment with Google, Inc., Mountain View, CA, in 2008, he developed a world-scale landmark recognition engine together with Google engineers, which has been highly praised and well publicized. He has served as a program committee member and reviewer of a number of prestigious international conferences and journals.

Sheng Tang received the Ph.D. degree in computer application technology from the Institute of Computing Technology, Chinese Academy of Sciences (ICT-CAS), Beijing, China, in 2006.

He is currently an Associate Professor with ICT-CAS. From 2006 to 2008, he was a TRECVID Team Leader with the Multimedia Computing Group, ICT-CAS. From February 2009 to February 2010, he was a Visiting Research Fellow with National University of Singapore, Singapore, under the instruction of Prof. C. Tat-Seng. His current research interests include pattern recognition and content-based multimedia retrieval and indexing.

Dr. Tang served as the reviewer for the *Journal of Visual Communication and Image Representation*, *Multimedia Tools and Applications*, and the *Journal of Computer Science and Technology*.

Xiangdong Zhou received the Ph.D. degree from Fudan University, Shanghai, China, in 2003.

He is currently an Associate Professor with the School of Computer Science and Technology, Fudan University. He spent one year as a Visiting Senior Research Fellow with the School of Computing, National University of Singapore, Singapore, in 2008. His current research interests include multimedia data management, analysis, and retrieval.

Yongdong Zhang received the Ph.D. degree in electronic engineering from Tianjin University, Tianjin, China, in 2002.

He has been a Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, since 2009. His current research interests include the field of video coding and transcoding, video analysis and retrieval, and universal media access.

Shouxun Lin received the Ph.D. degree from the Beijing University of Technology, Beijing, China, in 1998.

Since 1990, he has been an Associate Professor and became a Professor in 1995 with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing. His current research interests include multimedia processing and comparison, video coding, video analysis, multimedia indexing, statistical machine translation, and evaluation of computer–human interaction.

Tat-Seng Chua received the Ph.D. degree from the University of Leeds, Leeds, U.K.

He is currently a Professor with the School of Computing, National University of Singapore, Singapore. He was the Acting and Founding Dean of the School of Computing from 1998 to 2000. He spent three years as a Research Staff Member with the Institute of Systems Science (now I^2R), Beijing, China, in the late 1980s. He focuses on the use of relations between entities and external information and knowledge sources to enhance information processing. His current projects include news video retrieval and tracking, question–answering (QA), video QA, and information extraction on the web. His group participates regularly in TREC-QA and TRECVID news video retrieval evaluations. His current research interests include multimedia information processing, in particular, the extraction, retrieval, and QA of video and text information.