Action Recognition Using Multilevel Features and Latent Structural SVM

Xinxiao Wu, Dong Xu, *Member, IEEE,* Lixin Duan, Jiebo Luo, *Fellow, IEEE,* and Yunde Jia, *Member, IEEE*

***Abstract*—We first propose a new low-level visual feature, called spatio-temporal context distribution feature of interest points, to describe human actions. Each action video is expressed as a set of relative XYT coordinates between pairwise interest points in a local region. We learn a global Gaussian mixture model (GMM) (referred to as a universal background model) using the relative coordinate features from all the training videos, and then we represent each video as the normalized parameters of a video-specific GMM adapted from the global GMM. In order to capture the spatio-temporal relationships at different levels, multiple GMMs are utilized to describe the context distributions of interest points over multiscale local regions. Motivated by the observation that some actions share similar motion patterns, we additionally propose a novel mid-level class correlation feature to capture the semantic correlations between different action classes. Each input action video is represented by a set of decision values obtained from the pre-learned classifiers of all the action classes, with each decision value measuring the likelihood that the input video belongs to the corresponding action class. Moreover, human actions are often associated with some specific natural environments and also exhibit high correlation with particular scene classes. It is therefore beneficial to utilize the contextual scene information for action recognition. In this paper, we build the high-level co-occurrence relationship between action classes and scene classes to discover the mutual contextual constraints between action and scene. By treating the scene class label as a latent variable, we propose to use the latent structural SVM (LSSVM) model to jointly capture the compatibility between multilevel action features (e.g., low-level visual context distribution feature and the corresponding mid-level class correlation feature) and action classes, the compatibility between multilevel scene features (i.e., SIFT feature and the corresponding class correlation feature) and scene classes, and the contextual relationship between action classes and scene classes. Extensive experiments on UCF Sports, YouTube and UCF50 datasets demonstrate the effectiveness of the proposed multilevel features and action-scene interaction based LSSVM model for human action recognition. Moreover, our method generally achieves higher recognition accuracy than other state-of-the-art methods on these datasets.**

***Index Terms*—Action recognition, action-scene interaction, latent structural SVM, multilevel features.**

I. Introduction

RECOGNIZING human actions from videos still remains a challenging problem due to the large variations in human appearance, posture and body size within the same class. It also suffers from various factors such as cluttered background, occlusion, camera movement and illumination change. Many previous works can be roughly categorized into model-based methods and appearance-based approaches. Model-based methods [1], [2] usually rely on human body tracking or pose estimation in order to model the dynamics of individual body parts for action recognition. However, it is still a non-trivial task to accurately detect and track the body parts in unrestricted scenarios. Appearance-based approaches mainly employ appearance features for action recognition. For example, global space-time shape templates from image sequences are used in [3]–[5] to describe an action. However, in these methods, highly detailed silhouettes need to be extracted, which may be very difficult in a realistic video. Recently, approaches [6]–[8] based on local spatio-temporal interest points have shown much success in action recognition. Compared to the space-time shape and tracking based approaches, these methods do not require foreground segmentation or body parts tracking, so they are more robust to camera movement and low resolution. Each interest point is represented by its location (i.e., XYT coordinates) in the 3-D space-time volume and its spatio-temporal appearance information (e.g., the gray-level pixel values and 3DHoG). Using only the appearance information of interest points, many methods [6]–[8] model an action as a bag of independent and orderless visual words without considering the spatio-temporal contextual information of interest points.

In this paper, we first propose a new low-level visual feature by using multiple Gaussian mixture models (GMMs) to characterize the spatio-temporal context distributions about the relative coordinates between pairwise interest points over multiple space-time scales. Specifically, for each local region (i.e., subvolume) in a video, the relative coordinates between a pair of interest points in XYT space is considered as the spatio-temporal context feature. Then each action is represented by

Manuscript received February 15, 2012; revised June 29, 2012 and September 21, 2012; accepted December 12, 2012. Date of publication February 1, 2013; date of current version July 31, 2013. This work was supported in part by the Natural Science Foundation of China under Grant 61203274, 90920009. The work of D. Xu was supported in part by Multi-plAtform Game Innovation Centre (MAGIC) in Nanyang Technological University. MAGIC is funded by the Interactive Digital Media Programme Office (IDMPO) hosted by the Media Development Authority of Singapore. This paper was recommended by Associate Editor C. N. Taylor.

X. Wu and Y. Jia are with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, China (e-mail: wuxinxiao@bit.edu.cn; jiayunde@bit.edu.cn).

D. Xu is with the School of Computer Engineering, Nanyang Technological University, Singapore, 639798 (e-mail: dongxu@ntu.edu.sg).

L. Duan is with SAP Next Business and Technology, Singapore, 138602 (e-mail: lxduan@gmail.com).

J. Luo is with the Computer Science Department, University of Rochester, Rochester, NY 14627 USA (e-mail: jluo@cs.rochester.edu).

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.2013.2244794

a set of context features extracted from all pairs of interest points over all the local regions in a video volume. GMM is adopted to model the distribution of context features for each video. However, the context features from one video may not contain sufficient information to robustly estimate the parameters of GMM. Therefore, we first learn a global GMM [referred to as universal background model (UBM)] by using the context features from all the training videos and then describe each video as a video-specific GMM adapted from the global GMM via a maximum *a posteriori* (MAP) adaptation process. In this paper, multiple GMMs are exploited to cope with different levels of spatio-temporal contexts of interest points from different space-time scales.

Due to the biological and kinematic constraints in natural human movements, different action classes may have some similar motion patterns. For example, the actions of walking, jogging, and running share the typical motions of hands and legs, so it is beneficial to develop a descriptor for jogging by capturing the jogging–walking and jogging–running correlations. In our initial conference paper [9], we have developed a new learning method called multiple kernel learning with augmented features to learn an adapted classifier by leveraging the pre-learned classifiers of other action classes in order to exploit such correlations between different action classes. According to the resultant optimization problem [9], the decision values from these pre-learnt classifiers are used as features which can improve the recognition performance. In this paper, we therefore directly utilize a set of decision values as a mid-level class correlation feature in which each decision value is determined by the pre-learned classifier of the corresponding action class. The higher the correlation is, the larger the decision value becomes. Compared with the low-level visual feature, this decision value based feature describes higher level relationship among different action classes with more discriminative power and robustness.

Human actions are frequently associated with some specific natural environments, and exhibit co-occur relationship with particular scene classes. For example, the swing action often happens in a scene with a pool while it is common that the football action is with a football field. So if there is a football field within the scene, it is likely that there is a football action in the video. On the contrary, if there is a pool in the scene, the probability that the video belongs to the football action reduces. It is therefore natural and reasonable to utilize the scene information as a complementary clue for action recognition in the video. In this paper, we extract the bag of SIFT [10] descriptors from several frames randomly selected from the whole action video to describe the scene. Similarly, we also employ the mid-level class correlation feature of scene. The contextual relationship between action classes and scene classes is formulated by a contextual co-occurrence function of action labels and scene labels in an LSSVM model. Different from most of previous work [11], [12], we use LSSVM to jointly model the compatibility between multilevel action features and action classes, the compatibility between multilevel scene features and scene classes, and the contextual relationship between action classes and scene classes. In this paper, the scene label is treated as a latent variable and inferred implicitly during both learning and inference processes.

II. Related Work

A. Low-Level Spatio-Temporal Contextual Feature

Human action recognition from videos has attracted much attention in recent years [13] and how to extract discriminative and robust visual features has become an important issue. Many researchers have exploited the spatial and temporal context as another type of information for describing interest points. Kovashka and Grauman [14] exploited multiple bag-of-words models to represent the hierarchy of space-time configurations at different scales. Savarese *et al.* [15] used a local histogram to capture co-occurrences of interest points from the same visual word in a local region, and concatenated all the local histograms into a lengthy descriptor. Ryoo and Aggarwal [16] proposed a so-called feature×feature×relationship histogram to capture both appearance and relationship information between pairwise visual words. All these methods first utilize a vector quantization process to generate a codebook and adopt the bag-of-words representation. Then a contextual feature is designed to describe the spatio-temporal context information between visual words. In our work, we consider the relative XYT coordinates between pair-wise interest points, and the spatio-temporal context feature is directly extracted from the detected interest points rather than the visual words, which is modeled by a GMM. A global GMM represents the visual content of all the action classes including possible variations on human appearances, motion styles as well as environment conditions, and a video-specific GMM adapted from the global GMM provides additional information to distinguish the videos of different classes.

Bregonzio *et al.* [17] created clouds of interest points accumulated over multiple temporal scales, and extracted holistic features of the clouds as the spatio-temporal information of interest points. Zhang *et al.* [18] extracted motion words from motion images and utilized relative locations between the motion words and a reference point in a local region to establish the spatio-temporal context information. These methods above are based on some pre-processing steps such as human body detection and foreground segmentation. It will be shown that our method can still achieve promising performance even in complex environments with changing lighting and moving cameras without requiring any pre-processing steps. The methods in [19] and [20] proposed to use the distribution of pairwise relationships between edge primitivities to capture the shape of action in 2-D image. In contrast to [19] and [20], the relationship between the interest points in this paper is defined in 3-D video volume to capture both motion and shape of action. Moreover, they described the relational distribution just within the whole image, while we used multiple GMMs to model the context distribution of interest points at multiple space-time scales.

B. Mid-Level Semantic Feature

Related to our class correlation feature, several mid-level semantic features such as concept score [21] and attribute feature [22]–[24] have shown promising results for abstracting visual content and significantly improving the performance in visual recognition. Xu and Chang [21] proposed concept score features to characterize the semantic meaning of images for video event recognition. Liu *et al.* [23] explored both human-specified attribute and data-driven attribute classifiers to describe human actions by considering multiple semantic concepts. Farhadi *et al.* [22] used the L1-regularized logistic regression as the feature selection method to learn the object attributes that generalize well across different object categories. Parikh and Grauman [24] proposed relative attributes to capture more general semantic relationships which enable richer descriptions for images. Sadanand and Corso [25] used a number of action templates to calculate the matching values between the templates and an input action video. The matching results measuring the correlations between the templates and the input video are used to construct the final feature vector. These methods need to additionally collect huge labeled training data to learn the concept detectors, attribute classifiers, or action templates. In contrast, we only use the pre-learned classifiers from all the action classes to exploit the semantic relationship among different action classes. These pre-learned classifiers are obtained by using the same training data as that used for learning the subsequent LSSVM classifier, and we do not need to additionally collect any training data in our work. Moreover, when compared to the matching values from action templates in [25], the decision values from the SVM classifiers in our method are more discriminative to distinguish different actions.

C. Modeling the Scene and Action Interaction

Modeling of scene and action interaction has been explored in a few recent papers. Ikizler-Cinbis and Sclaroff [12] extracted multiple features of people, objects and scene, and then combined these features for classification using linear SVM and multiple kernel learning methods in a multiple instance learning framework. Different from their work, we explicitly exploit the action-scene interaction by integrating the contextual co-occurrence relationship into the LSSVM model. Marszalek *et al.* [11] first used a text-mining approach to discover the conditional probability matrix encoding the relations between action and scene from movie script, and then linearly combined the separately learned action classifier and scene classifier according to the conditional probability. In our work, the action classifier, the scene classifier and the interaction between action and scene are jointly modeled in a unified framework. Therefore, the action and scene classifiers as well as action-scene interaction function are simultaneously learned during the optimization process. Moreover, we treat the scene label as a latent variable and do not require the ground truth of the scene label in the training data.

III. Multilevel Features for Action Description

A. Low-Level Spatio-Temporal Context Distribution Feature

We use the interest point detector proposed by Dollar *et al.* [6] which respectively employs 2-D Gaussian filter in the spatial direction and 1-D Gabor filter in the temporal direction. The two separate filters can produce high response at points with significant spatio-temporal intensity variations. Fig. 1 shows some examples of detected interest points of human actions. The subfigures on the left column represent the detected interest points (depicted by red squares) in each frame from a video and the subfigures on the right column show all the interest points (depicted by red dots) from a video accumulated in one single image. It is worth mentioning that most detected interest points are near the body parts that have major contribution to the action classification.

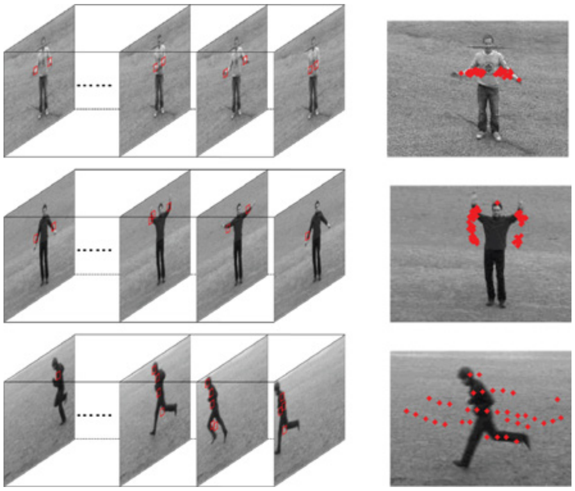


Fig. 1. Samples of detected interest points of handclapping, handwaving, and running.

1) *Multiscale Spatio-Temporal Context Extraction:* As an important type of action representation, the spatio-temporal context information of interest points characterizes both spatial relative layout of human body parts and temporal evolution of human poses. In order to represent the spatio-temporal context between interest points, we propose a new local spatio-temporal context feature using a set of XYT relative coordinates between any pairs of interest points in a local region. Suppose there are R interest points in a local region, then the number of pairwise relative coordinates is $R(R-1)$. For efficient computation and compact description, we define a center interest point $[X_c \quad Y_c \quad T_c]^T$ as the mean value of all the coordinates $[X_i \quad Y_i \quad T_i]^T$, where $[X_i \quad Y_i \quad T_i]^T$ represent the ith interest point in a local video region. Consequently, the spatio-temporal contextual information of interest points is characterized by R relative coordinates between all the interest points and the center interest point, i.e., $s_i = [X_i - X_c \quad Y_i - Y_c \quad T_i - T_c]^T$, $i = 1, 2, ..., R$. As shown in Fig. 2(b), the red star represents the center interest point of all interest points in a local region. The relative coordinates are normalized by the mean of distances between all the interest points and the center interest point. A large number of relative coordinates extracted from all the local regions over the entire video collectively describe the spatio-temporal context information of interest points for an action.

To capture the spatio-temporal context of interest points at different space-time scales, we use multiscale local regions across multiple space-time scales to generate multiple sets of local context features (i.e., XYT relative coordinates). Each

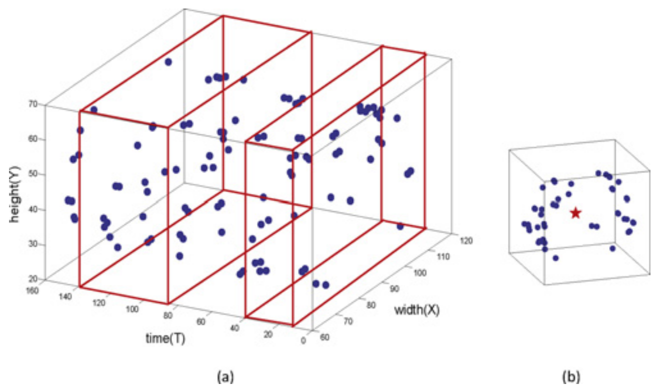


Fig. 2. Multiscale spatio-temporal context extraction from interest points of one handwaving video in (a) 3-D video volume and (b) local region.

set represents the spatio-temporal context at one space-time scale. Fig. 2(a) illustrates the detected interest points of one video from the handwaving action, where the blue dots are the interest points in 3-D video volume and the red bounding boxes represent certain local regions at different space-time scales. In our experiments, for computational simplicity and efficiency, we set the spatial size of the local region the same as that of each frame, and we use multiple temporal scales represented by different numbers of frames. Consequently, the local regions are generated by simply moving a spatio-temporal window frame by frame through the video. Suppose there are T local regions at one scale and R relative coordinates in each local region, then the total number of relative coordinates in the entire video at this scale is $N = RT$.

2) *Multiscale Spatio-Temporal Context Distribution Feature:* For each action video, a video-specific GMM is employed to characterize the distribution of the set of spatio-temporal context features at one space-time scale. Compared with the conventional bag-of-words framework, the GMM is a compact description of the underlying distribution of all spatio-temporal context features within an action video, which is more robust to noise because of the effective two-step distribution estimate process discussed below. Considering the spatio-temporal context features extracted from one video may not contain sufficient information to robustly estimate the parameters of the video-specific GMM, we therefore propose a two-step approach. We first train a global GMM (also referred to as UBM) using all the spatio-temporal context features from all the training videos. The global GMM can be represented as its parameter set $\{(m_k, \mu_k, \Sigma_k)|_{k=1}^K\}$ where K is the total number of GMM components. m_k, μ_k and Σ_k are the weight, the mean vector and the covariance matrix of the kth Gaussian component, respectively. Note we have the constraint $\sum_{k=1}^K m_k = 1$. As suggested in [26], the covariance matrix Σ_k is set to be a diagonal matrix for computational efficiency. We adopt the well known expectation-maximization algorithm to iteratively update the weight, the mean and the covariance matrix. m_k is initialized to the uniform weights. We partition all the training context features into K clusters and use the samples in each cluster to initialize μ_k and Σ_k.

The video-specific GMM for each video can be generated from the global GMM via an MAP adaption process. Given the set of spatio-temporal context features $\{s_1, s_2, ..., s_N\}$ extracted from an action video V where $s_i \in \mathbb{R}^3$ denotes the ith context feature vector (i.e., XYT relative coordinates), we introduce an intermediate variable $\eta(k|s_i)$ to indicate the membership probability of s_i belonging to the kth GMM component:

$$\eta(k|s_i) = \frac{m_k P_k(s_i|\theta_k)}{\sum_{j=1}^K m_j P_j(s_i|\theta_j)},$$

where $P_k(s_i|\theta_k)$ represents the Gaussian probability density function with $\theta_k = \{\mu_k, \Sigma_k\}$. Note we have the constraint $\sum_{k=1}^K \eta(k|s_i) = 1$. Let $\zeta_k = \sum_{i=1}^N \eta(k|s_i)$ be the soft count of all the context features $s_i|_{i=1}^N$ belonging to the kth GMM component. Then, the kth component of the video-specific GMM of any video V can be adapted as follows:

$$\bar{\mu}_k = \frac{\sum_{i=1}^N \eta(k|s_i)s_i}{\zeta_k}, \hat{\mu}_k = (1-\rho_k)\mu_k + \rho_k\bar{\mu}_k, \hat{m}_k = \frac{\zeta_k}{N},$$

where $\bar{\mu}_k$ is the expected mean of the kth component based on the training samples $s_i|_{i=1}^N$ and $\hat{\mu}_k$ is the adapted mean of the kth component. The weighting coefficient $\rho_k = \frac{\zeta_k}{\zeta_k + r}$ is introduced to improve the estimation accuracy of the mean vector. Following [26], only the means and weighs of GMM are adapted to better cope with the instability problem during the parameter estimation and reduce the computational cost. Therefore, the video V is represented by the video-specific GMM parameter set

$$\{(\hat{m}_k, \hat{\mu}_k, \Sigma_k)|_{k=1}^K\},$$

where Σ_k is the covariance matrix of the kth Gaussian component from UBM. Finally, the spatio-temporal context distribution feature x of video V is represented by

$$x = [v_1^T, v_2^T, ..., v_K^T]^T \in \mathbb{R}^D, v_k = \sqrt{\frac{\hat{m}_k}{2}}\Sigma_k^{-\frac{1}{2}}\hat{\mu}_k \in \mathbb{R}^3,$$

where $D = 3K$ is the feature dimension of x. To capture the context distributions at different spatio-temporal levels, we propose to use multiple GMMs with each GMM representing the context distribution of interest points at one space-time scale. The previous work about multiscale spatio-temporal context distribution feature has been published in [9].

B. Other Low-Level Visual Features

To complement the spatio-temporal context distribution information of the interest points, we extract the appearance information from the cuboids around the interest points. Specially, we first normalize the gray-level pixel values in each cuboid and then flatten the normalized cuboid into a vector which is further reduced via principle component analysis by preserving 98% energy. Finally, the standard bag-of-words model is adopted to generate the appearance feature vector of interest points. Moveover, three dense trajectory features (i.e., trajectory, HOG, MBH) proposed by Wang *et al.* [27] are also combined to further improve the recognition performance. Consequently, we use five types of heterogeneous and complementary low-level visual features (i.e., spatio-temporal context distribution and appearance features of interest points, and three types of dense trajectory features) to describe the action from video.

C. Mid-Level Class Correlation Feature

The extracted low-level visual features only represent the visual information of action video and their discriminative capability is limited. Thus we newly propose a higher level feature, called class correlation feature which captures the correlations between different action classes, to abstract the visual content of video. The intuitive explanation is: some different action classes may often share certain similar motion patterns of human body parts and such class correlation between different action classes can be used to distinguish different actions [28].

For each action video, its mid-level class correlation feature is represented by a set of decision values from all the pre-learned classifiers, which can represent the semantic meaning of the action video to some extent. The pre-learned classifiers can be trained using any classification algorithm, such as SVMs, Naive Bayes, decision tree, etc. In this paper, we employ the SVM classifiers. Specifically, using each type of low-level visual feature, an independent SVM classifier is trained for each action class. Based on five types of low-level visual features mentioned in Section III-B, five independent SVMs are learned for each action class to produce the decision values. Let us denote $f_c^l(x)$ as the pre-learned classifier of the cth action class from the lth type of visual feature and x as the action video. Using the lth type of visual feature, the likelihood that the video x belongs to the cth action class is modeled by the classification score $h^l = f_c^l(x)$, and the corresponding class correlation feature of x for the lth type of visual feature is then represented by $H^l = [h_1^l, h_2^l, ..., h_C^l]^T \in \mathbb{R}^C$, where C is the number of action classes. Finally, the mid-level class correlation feature of x is constructed by simply concatenating H^l from all types of visual features.

IV. Modeling Action and Scene Interaction Using LSSVM

A. Model Formulation

Let $x \in \mathcal{X}$ be the input action video and $y \in \mathcal{Y}$ be the output action label, we aim to learn a discriminative function $f_{\mathbf{w}}(x, y) = \mathbf{w}^T\Phi(x, y) : \mathcal{X} \times \mathcal{Y} \rightarrow \mathbf{R}$ over input–output pairs, where $\Phi(x, y)$ is a joint feature vector that describes the relationship between x and y, and $\mathbf{w}$ is the parameter vector. By maximizing $f_{\mathbf{w}}(x, y)$ over all y for any given input x, we can derive a prediction $y^* = \arg\max_{y\in\mathcal{Y}} f_{\mathbf{w}}(x, y)$.

However, in this paper, the input–output relationship is not completely characterized by the pair of action video and class label $(x, y) \in \mathcal{X} \times \mathcal{Y}$, because it also depends on the unobserved latent scene class label $s \in \mathcal{S}$. Hence, we extend the joint feature vector $\Phi(x, y)$ to $\Phi(x, y, s)$ to describe the relation among the input action video x, the output action class label y, and the latent scene class label s. Accordingly, the prediction problem takes the following form: $f_{\mathbf{w}}(x) = \max_{y\in\mathcal{Y}} \max_{s\in\mathcal{S}} f_{\mathbf{w}}(x, y, s) = \max_{y\in\mathcal{Y}} \max_{s\in\mathcal{S}} \mathbf{w}^T\Phi(x, y, s)$. The model parameter vector $\mathbf{w}$ has three parts $\mathbf{w} = \{\mathbf{w}_a; \mathbf{w}_s; \mathbf{w}_{as}\}$ and $\mathbf{w}^T\Phi(x, y, s)$ is defined as

$$\mathbf{w}^T\Phi(x, y, s) = \mathbf{w}_a^T\phi_a(x, y) + \mathbf{w}_s^T\phi_s(x, s) + \mathbf{w}_{as}^T\phi_{as}(y, s).$$

1) *Action Class Model* $\mathbf{w}_a^T\phi_a(x, y)$*:* This potential function measures the compatibility between an action video x and an action label y, defined by $\sum_{i=1}^{C_a}\sum_{j=1}^{L_a} \mathbf{w}_{ij}^{av\,T}\varphi_j^{av}(x)I_i(y) + \sum_{i=1}^{C_a}\sum_{j=1}^{L_a} \mathbf{w}_{ij}^{ac\,T}\varphi_j^{ac}(x)I_i(y)$. C_a and L_a respectively denote the number of action classes and the number of action feature types. $\varphi_j^{av}(x) \in \mathbf{R}^{d_j}$ represents the jth type of action visual feature extracted from the video x with the vector dimension of d_j, and $\varphi_j^{ac}(x) \in \mathbf{R}^{C_a}$ represents the jth type of action class correlation feature generated by the pre-learned classifiers of the jth type of action visual feature with the vector dimension of C_a. The parameters $\mathbf{w}_{ij}^{av} \in \mathbf{R}^{d_j}$ and $\mathbf{w}_{ij}^{ac} \in \mathbf{R}^{C_a}$ are respectively the weight vectors for the features $\varphi_j^{av}(x)$ and $\varphi_j^{ac}(x)$. $I_i(y)$ is an indicator function, namely, $I_i(y) = 1$ if $y = i$, and $I_i(y) = 0$ otherwise. $\sum_{i=1}^{C_a}\sum_{j=1}^{L_a} \mathbf{w}_{ij}^{av\,T}\varphi_j^{av}(x)I_i(y)$ represents the compatibility between the low-level action visual features and the action label, and $\sum_{i=1}^{C_a}\sum_{j=1}^{L_a} \mathbf{w}_{ij}^{ac\,T}\varphi_j^{ac}(x)I_i(y)$ indicates the compatibility between the mid-level action class correlation features and the action label.

2) *Scene Class Model* $\mathbf{w}_s^T\phi_s(x, s)$*:* This potential function models the compatibility between an action video x and a scene label s, defined by $\sum_{i=1}^{C_s}\sum_{j=1}^{L_s} \mathbf{w}_{ij}^{sv\,T}\varphi_j^{sv}(x)I_i(s) + \sum_{i=1}^{C_s}\sum_{j=1}^{L_s} \mathbf{w}_{ij}^{sc\,T}\varphi_j^{sc}(x)I_i(s)$. L_s and C_s denote the number of scene features and the number of scene classes, respectively. $\varphi_j^{sv}(x) \in \mathbf{R}^{e_j}$ indicates the jth type of scene visual feature with the vector dimension of e_j, and $\mathbf{w}_{ij}^{sv} \in \mathbf{R}^{e_j}$ is the weight vector for the feature $\varphi_j^{sv}(x)$. $\varphi_j^{sc}(x) \in \mathbf{R}^{C_s}$ is the jth type of scene class correlation feature and $\mathbf{w}_{ij}^{sc} \in \mathbf{R}^{C_s}$ is the weight vector for the feature $\varphi_j^{sc}(x)$. $\sum_{i=1}^{C_s}\sum_{j=1}^{L_s} \mathbf{w}_{ij}^{sv\,T}\varphi_j^{sv}(x)I_i(s)$ is the compatibility between the low-level scene visual features and the scene label, while $\sum_{i=1}^{C_s}\sum_{j=1}^{L_s} \mathbf{w}_{ij}^{sc\,T}\varphi_j^{sc}(x)I_i(s)$ indicates the compatibility between the mid-level class correlation features and the scene label. In our work, we extract the local SIFT features from randomly selected frames in the video and use the bag-of-words model.

3) *Action-Scene Interaction Model* $\mathbf{w}_{as}^T\phi_{as}(y, s)$*:* This potential function represents the contextual co-occurrence relation between an action label y and a scene label s. It is parameterized as $\mathbf{w}_{as}^T\phi_{as}(y, s) = \sum_{i=1}^{C_a}\sum_{j=1}^{C_s} \mathbf{w}_{ij}^{as}I_i(y) \cdot I_j(s)$, where $\mathbf{w}_{ij}^{as}$ indicates how likely the action class is i and the scene class is j. For example, suppose the ith action class is swimming action and the jth scene class indicates a pool, it is likely that $\mathbf{w}_{ij}^{as}$ has a larger value since the swimming action always happens in the scene with a pool.

B. Inference

Given the model parameter $\mathbf{w} = \{\mathbf{w}_a; \mathbf{w}_s; \mathbf{w}_{as}\}$, the inference problem is to find the optimal action label y^* for a test video x and we need to solve the following optimization problem: $(y^*, s^*) = \arg\max_{y,s} \mathbf{w}^T\Phi(x, y, s)$. For simplicity, we can directly enumerate all the possible action-scene label pairs (y, s) to predict the optimal action label y^* for x. The values of y and s are respectively set from 1 to C_a and from 1 to C_s.

C. Training

Given a set of N training examples $\{(x_n, y_n)\}$, $n = 1, 2, ..., N$, our goal is to learn the model parameter $\mathbf{w}$. Since the scene label s is unobserved and treated as a latent variable during the training process, we adopt the Latent Structural SVM framework [29], [30] by formulating the following optimization problem:

$$\begin{aligned} \min_{\mathbf{w},\xi_n} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_n \xi_n \\ \text{s.t.} \quad & \max_s \mathbf{w}^T\Phi(x_n, y_n, s) - \max_s \mathbf{w}^T\Phi(x_n, y, s) \\ & \geq \Delta(y_n, y) - \xi_n, \xi_n \geq 0, \forall n \quad \forall y \end{aligned} \tag{1}$$

where $\Delta(y_n, y)$ is a loss function measuring the cost incurred by predicting the ground-truth label y_n as y. In this paper, we use a simple 0–1 loss as $\Delta(y_n, y) = 1$ if $y_n \neq y$, and $\Delta(y_n, y) = 0$ otherwise. The constraint in (1) can be explained as follows: for the nth training sample, the score $\max_s \mathbf{w}^T\Phi(x_n, y_n, s)$ associated with the ground-truth class label y_n should be no less than the score $\max_s \mathbf{w}^T\Phi(x_n, y, s)$ associated with any hypothesized class label y. We rewrite the constrained optimization problem in (1) as an unconstrained problem

$$\min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_n (L_n - R_n) \tag{2}$$

where $L_n = \max_{y,s}(\Delta(y_n, y) + \mathbf{w}^T\Phi(x_n, y, s))$ and $R_n = \max_s \mathbf{w}^T\Phi(x_n, y_n, s)$. On the other hand, it is easy to show that (2) can be rewritten as (1), if we define $\xi_n = L_n - R_n$. Therefore, (1) and (2) are equivalent. We employ the non-convex bundle optimization technique in [31] to solve (2). Specifically, this optimization algorithm iteratively builds an increasingly accurate piecewise quadratic approximation of (2) and converges to an optimal solution of $\mathbf{w}$, which requires the calculation of the subgradient of $L_n - R_n$. Suppose $(y^\star, s^\star) = \arg\max_{y,s}(\Delta(y_n, y) + \mathbf{w}^T\Phi(x_n, y, s))$ and $s^\dagger = \arg\max_s \mathbf{w}^T\Phi(x_n, y_n, s)$, then $\partial_{\mathbf{w}}(L_n - R_n)$ can be calculated as $\Phi(x_n, y^\star, s^\star) - \Phi(x_n, y_n, s^\dagger)$. Note that the inferences on $\max_{y,s}(\Delta(y_n, y) + \mathbf{w}^T\Phi(x_n, y, s))$ and $\max_s \mathbf{w}^T\Phi(x_n, y_n, s)$ can be respectively solved via the enumeration of all the possible label pairs (y, s) and the enumeration of all the possible scene labels s. In our work, we initialize the latent scene class labels as the same as the action class labels.

V. Experiments

A. *Human Action Datasets*

The UCF Sports dataset [32] contains ten different types of sports actions. The dataset consists of 149 real videos with large intra-class variabilities. Each action class is performed in different number of ways, and the frequencies of various actions also differ considerably. In order to increase the amount of training samples, we extend the dataset by adding a horizontally flipped version of each video sequence to the dataset as suggested in [33]. In the leave-one-sample-out cross validation setting, one original video sequence is used as the test data while the rest original video sequences together with their flipped versions are employed as the training data. Following [33], the flipped version of the test video sequence is not included in the training set.

The YouTube dataset [34] contains 11 action classes. There are in total 1168 videos and the videos for each class are divided into 25 related groups with each group consisting of 4 or more than 4 videos. Following the evaluation method in [34], the leave-one-group-out cross validation setting is applied over those groups.

The UCF50 dataset[1] is the largest and most difficult dataset for action recognition. It contains 50 different actions. Each action class consists of 25 to 30 related action groups and each group contains 4 to 23 videos. In our experiment, we just use the first 25 groups of each action class and adopt the leave-one-group-out cross validation over these groups similar to the YouTube dataset, as suggested in [34]. Following [25], we also adopt the 10-fold video-wise cross validation and 5-fold group-wise cross validation to evaluate the proposed method.

B. *Experimental Setting*

For interest points detection, the spatial and temporal scale parameters σ and τ are empirically set by $\sigma = 2$ and $\tau = 2.5$, respectively. The size of cuboid is empirically fixed as $7 \times 7 \times 5$ and 1000 interest points are extracted from each video. For the spatio-temporal (ST) context distribution feature of interest points, the number of space-time scales is fixed to five and the number of Gaussian components in GMM (i.e., K) is set to 2000. These parameters empirically lead to good results for a wide range of datasets. Similarly as in the existing work [27], [33] using the bag-of-words model, we randomly select 100 000 ST context features (i.e., XYT relative coordinates) to train the global GMM. For the appearance feature of interest points, the number of appearance words is set to 2000. For the SIFT feature of scene, SIFT descriptors are extracted from the 20% frames randomly selected from each video and the number of SIFT words is fixed to 2000. Three descriptors [27] (i.e., trajectory, HOG, MBH) of dense trajectory are additionally extracted as complementary action visual features to improve the recognition performance. Following [27], we use the bag-of-features approach and the number of visual words per descriptor is fixed to 4000. Therefore, the proposed ST context feature, the appearance feature, and three dense trajectory features are used as the action visual features. The SIFT feature represents the scene visual feature. The parameter C in LSSVM is fixed as the default value (i.e., $C = 1$) as in SVM.

C. *Experimental Results*

1) *Recognition Results Using Different Features:* Based on the ST context features from multiple space-time scales, Fig. 3 compares the bag-of-words model and our GMMs with different number of cluster centers for the complex YouTube dataset. The horizontal axis and vertical axis indicate the number of cluster centers and recognition accuracy, respectively. The number of space-time scales is fixed to five. It is obvious that the results using our proposed GMMs are better than those using the bag-of-words model, which demonstrates the effectiveness of the GMMs.

[1] Available at http://server.cs.ucf.edu/˜vision/data/UCF50.rar.

TABLE I
ACCURACIES USING DIFFERENT FEATURES ON THREE DATASETS

	UCF Sports		YouTube		UCF50	
Feature	Vis.	Vis.+Corr.	Vis.	Vis.+Corr.	Vis.	Vis.+Corr.
Trajectory	61.55%	57.93%	51.90%	54.50%	45.12%	50.62%
HOG	78.54%	78.82%	63.31%	65.70%	54.71%	57.48%
MBH	76.52%	83.84%	73.31%	78.15%	66.67%	71.77%
Appearance	75.63%	80.25%	54.68%	57.62%	47.73%	51.53%
ST Context Distribution	64.49%	68.99%	39.79%	44.84%	36.33%	39.05%
Action without ST Context Distribution	86.11%	88.76%	82.64%	83.17%	77.75%	81.34%
Action	89.34%	89.93%	82.31%	84.44%	80.31%	83.86%
Scene	70.04%	73.50%	54.48%	55.28%	41.67%	43.59%
Action+Scene	90.11%	91.98%	82.62%	85.97%	81.11%	84.77%

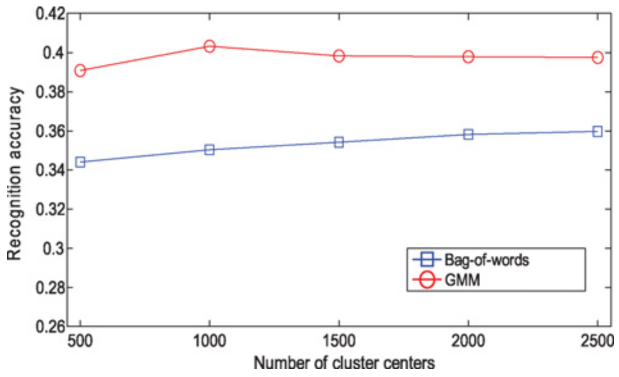


Fig. 3. Comparison between the bag-of-words model and GMMs based on the ST context features on the YouTube dataset.

TABLE II
ACCURACIES OF DIFFERENT METHODS ON THE UCF SPORTS DATASET

Methods	Recognition Accuracy
SVM	91.98%
LSSVM (our method)	92.48%
Kovashka and Grauman [14]	87.27%
Wang *et al.* [33]	85.6%
Wang *et al.* [27]	88.2%
Le *et al.* [35]	86.5%
Shabani *et al.* [36]	91.5%
Sadanand and Corso [25]	95.0%
Wu *et al.* [9]	91.3%
Yeffet and Wolf [37]	79.3%
Wu *et al.* [38]	89.7%
Rodriguez *et al.* [32]	69.2%
Bregonzio *et al.* [39]	86.9%

TABLE III
ACCURACIES OF DIFFERENT METHODS ON THE YOUTUBE DATASET

Methods	Recognition Accuracy
SVM	85.97%
LSSVM (our method)	87.01%
Liu *et al.* [34]	71.2%
Ikizler-Cinbis *et al.* [12]	75.21%
Wang *et al.* [27]	84.2%
Le *et al.* [35]	75.8%
Bregonzio *et al.* [39]	64.0%
Chakraborty *et al.* [40]	86.98%

Table I lists the recognition results using different features on the datasets of UCF Sports, YouTube and UCF50, respectively. The first three rows called "Trajectory," "HOG," and "MBH" respectively indicate the trajectory, HOG and MBH features of dense trajectory. The next two rows called "Appearance" and "ST Context Distribution" respectively represent the appearance and ST context distribution feature of interest points. The last four rows called "Action without ST Context Distribution," "Action," "Scene" and "Action+Scene" respectively indicate all the action features excluding the ST context distribution feature (i.e., appearance, trajectory, HOG and MBH), all the five types of action features (i.e., ST context, appearance, trajectory, HOG and MBH), the SIFT feature of scene, and the combination of all the five types of action features and scene feature. The column of "vis." represents the results using the low-level visual features while the "vis.+corr." column represents the results using the combination of visual features and the corresponding class correlation features.

From Table I, we have the observations as follows.

1) While the ST context distribution feature itself is less effective than the combination of other four types of action features, the combination of ST context distribution feature with other action features can generally improve the recognition performance by additionally characterizing the "where" property of interest points (see the results from the fifth row to the seventh row).
2) Although the scene feature is not as effective as the action features, the combination of the dynamic action features and static scene feature outperforms the action features or scene feature (see the results from the last three rows).
3) The combination of low-level visual features and mid-level class correlation features achieves the best results in almost all the cases, which demonstrates the effectiveness of using the decision values from the pre-learned classifiers of all the action classes to improve the recognition performance.

2) *Results From Latent Structural SVM Using Multilevel Features:* Tables II–IV compare the linear SVM and the proposed LSSVM using both the low-level visual and mid-level class correlation features on the three datasets. It is obvious that LSSVM outperforms SVM in terms of recognition accuracy on all the datasets, which clearly demonstrates that

TABLE IV

ACCURACIES OF DIFFERENT METHODS ON THE UCF50 DATASET

Methods	Leave-one-group-out	10 fold-video-wise	5 fold-group-wise
SVM	84.77%	86.91%	82.09%
LSSVM (our method)	85.87%	88.04%	83.02%
Sadanand and Corso [25]	–	76.4%	57.9%
Oliva and Torralba [41]	–	–	38.8%
Laptev *et al.* [33], [42]	–	–	47.9%

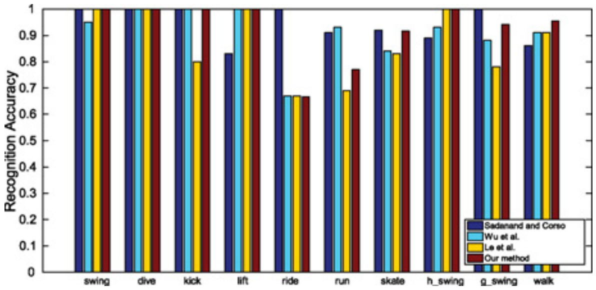


Fig. 4. Comparison of recognition accuracy per action class between our method and the work respectively from Sadanand and Corso [25], Wu *et al.* [9] and Le *et al.*'s [35] work on the UCF Sports dataset.

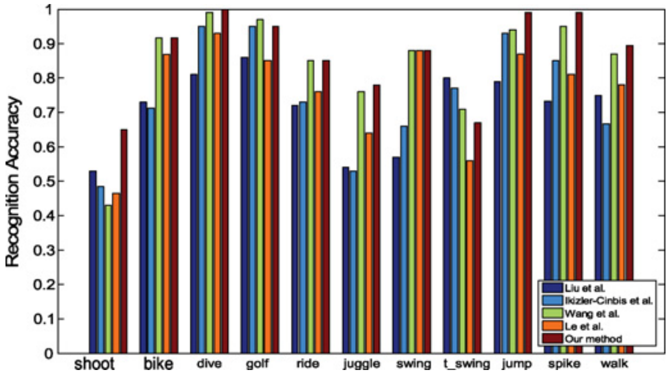


Fig. 5. Comparison of recognition accuracy per action class between our method and the work respectively from [34], [12], [27] and [35] on the YouTube dataset.

it is helpful to exploit the contextual co-occurrence between actions and scenes for distinguishing different actions.

In Table II, we report the recognition accuracies of the state-of-the-art methods on UCF Sports dataset. While some methods [32], [37], [38], [39] use a subset of videos or different cross validation settings, the most recent methods [9], [14], [25], [27], [33], [35], [36] employ the same leave-one-sample-out cross validation setting. So in this paper we strictly follow the experimental setting in the latest methods [9], [14], [25], [27], [33], [35], [36]. The results clearly show that our method outperforms most of the state-of-the-art methods excluding [25]. In [25], a number of action templates are additionally collected from the labeled action datasets (i.e., UCF50 and KTH datasets) in order to generate the final feature vector, which can significantly improve the recognition performance. In our work, we do not use any additional datasets and still achieve promising result on the UCF sports dataset. In Table III, we report the recognition accuracies of the state-of-the-art methods which adopt the same leave-one-group-out cross validation strategy on the YouTube dataset and our method achieves the best result. In Table IV, we show the recognition results on the UCF50 dataset using leave-one-group-out, 5-fold-group-wise and 10-fold-video-wise cross validation strategies. It is obvious that our method outperforms the state-of-the-art methods on this most complex and challenging action dataset with 50 action classes and large intra-class variations.

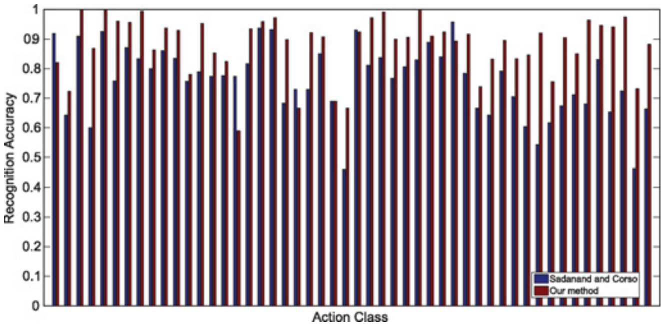


Fig. 6. Comparison of recognition accuracy per action class between the work in [25] and our method on the UCF50 dataset. The vertical axis represents the recognition accuracy and the horizontal axis indicates different action classes.

Fig. 4 illustrates the recognition accuracies per action class among our method and [9], [25], and [35] on the UCF Sports dataset. We also compare the results for each action class among our method and [12], [27], [34], and [35] on the YouTube dataset in Fig. 5. The comparison of recognition accuracies per action class between our method and [25] on the UCF50 dataset is demonstrated in Fig. 6. For most action classes, our method achieves the best or comparable results when compared with other methods.

VI. CONCLUSION

We have proposed a new low-level visual feature by using multiple GMMs to represent the distributions of local spatio-temporal context between interest points at different space-time scales that describe the "where" property of interest points. To exploit the correlation between different action classes, a novel mid-level class correlation feature is presented by using the decision values from prelearned classifiers of all the action classes. We also propose to use the latent structural SVM for jointly modeling the compatibility between multilevel action features and action labels, the compatibility between multilevel scene features and scene labels, and the contextual relationship between action labels and scene labels. Extensive experiments on UCF Sports, YouTube, and UCF50 datasets demonstrate that our method outperforms the state-of-the-art algorithms for action recognition.

In the future, we plan to extend the latent structural SVM to multiple kernel learning methods for effectively fusing multilevel features as well as to explore more constraints between actions and scenes for improving action recognition results. We will also apply our proposed method to other applications, such as action detection from continuous videos.

REFERENCES

[1] C. Fanti, L. Zelnik-manor, and P. Perona, “Hybrid models for human motion recognition,” in *Proc. IEEE CVPR*, pp. 1166–1173 Jun. 2005.

[2] A. Yilmaz, “Recognizing human actions in videos acquired by uncalibrated moving cameras,” in *Proc. IEEE ICCV*, pp. 150–157 Oct. 2005.

[3] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, “Actions as space-time shapes,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 12, pp. 2247–2253, Dec. 2007.

[4] A. Yilmaz and M. Shah, “Actions sketch: A novel action representation,” in *Proc. IEEE CVPR*, pp. 984–989, Jun. 2005.

[5] X. Wu, Y. Jia, and W. Liang, “Incremental discriminant-analysis of canonical correlations for action recognition,” *Pattern Recognit.*, vol. 43, no. 12, pp. 4190–4197, Dec. 2010.

[6] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proc. IEEE Int. Workshop PETS*, pp. 65–72 Jun. 2005.

[7] C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: A local SVM approach,” in *Proc. ICPR*, pp. 32–36, Aug. 2004.

[8] J. C. Niebles, H. Wang, and L. Fei-fei, “Unsupervised learning of human action categories using spatial-temporal words,” *Int. J. Comput. Vision*, vol. 79, no. 3, pp. 299–318, Sep. 2008.

[9] X. Wu, D. Xu, L. Duan, and J. Luo, “Action recognition using context and appearance distribution features,” in *Proc. IEEE CVPR*, pp. 489–496, Jun. 2011.

[10] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[11] M. Marszaek, I. Laptev, and C. Schmid, “Actions in context,” in *Proc. IEEE CVPR*, pp. 2929–2936, Jun. 2009.

[12] N. Ikizler-Cinbis and S. Sclaroff, “Object, scene and actions: Combining multiple features for human action recognition,” in *Proc. ECCV*, pp. 494–507, Sep. 2010.

[13] J. Aggarwal and M. Ryoo, “Human activity analysis: A review,” *ACM Comput. Surveys*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[14] A. Kovashka and K. Grauman, “Learning a hierarchy of discriminative space-time neighborhood features for human action recognition,” in *Proc. IEEE CVPR*, pp. 2046–2053, Jun. 2010.

[15] S. Savarese, A. Delpozo, J. C. Niebles, and L. Fei-fei, “Spatial-temporal correlatons for unsupervised action classification,” in *IEEE Workshop on Motion and video Computing*, pp. 1–8, Jan. 2008.

[16] M. Ryoo and J. Aggarwal, “Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities,” in *Proc. IEEE ICCV*, pp. 1593–1600, Sep. 2009.

[17] M. Bregonzio, S. Gong, and T. Xiang, “Recognising action as clouds of space-time interest points,” in *Proc. IEEE CVPR*, pp. 1984–1955, Jun. 2009.

[18] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, “Motion context: A new representation for human action recognition,” in *Proc. ECCV*, pp. 817–829, Oct. 2008.

[19] I. R. Vega and S. Sarkar, “Statistical motion model based on the change of feature relationships: Human gait-based recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 10, pp. 1323–1328, Oct. 2003.

[20] S. Nayak, S. Sarkar, and B. L. Loeding, “Distribution-based dimensionality reduction applied to articulated motion recognition,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 795–810, May 2009.

[21] D. Xu and S.-F. Chang, “Video event recognition using kernel methods with multilevel temporal alignment,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 1985–1997, Nov. 2008.

[22] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth, “Describing objects by their attributes,” in *Proc. IEEE CVPR*, pp. 1778–1785, Jun. 2009.

[23] J. Liu, B. Kuipers, and S. Savarese, “Recognizing human actions by attributes,” in *Proc. IEEE CVPR*, pp. 3337–3344, Jun. 2011.

[24] D. Parikh and K. Grauman, “Relative attributes,” in *Proc. IEEE ICCV*, pp. 503–510, Nov. 2011.

[25] S. Sadanand and J. J. Corso, “Action bank: A high-level representation of activity in video,” in *Proc. IEEE CVPR*, pp. 1234–1241, Jun. 2012.

[26] D. Xu, Y. Huang, Z. Zeng, and X. Xu, “Human gait recognition using patch distribution feature and locality-constrained group sparse representation,” *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 316–326, Jan. 2012.

[27] H. Wang, A. Klaer, C. Schmid, and C.-L. Liu, “Action recognition by dense trajectories,” in *Proc. IEEE CVPR*, pp. 3169–3176, Jun. 2011.

[28] L. Duan, D. Xu, I. Tsang, and J. Luo, “Visual event recognition in videos by learning from web data,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 9, pp. 1667–1680, Sep. 2012.

[29] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, “Object detection with discriminatively trained part-based models.” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.

[30] C.-N. J. Yu and T. Joachims, “Learning structural svms with latent variables,” in *Proc. ICML*, pp. 1169–1176, Jun. 2009.

[31] T. A. T.-M.-T. Do, “Large margin training for hidden markov models with partially observed states,” in *Proc. ICML*, pp. 265–272, Jun. 2009.

[32] M. D. Rodriguez, J. Ahmed, and M. Shah, “Action mach a spatio-temporal maximum average correlation height filter for action recognition,” in *Proc. IEEE CVPR*, pp. 1–8, Jun. 2008.

[33] H. Wang, M. M. Ullah, A. Klaer, I. Laptev, and C. Schmid, “Evaluation of local spatio-temporal features for action recognition,” in *Proc. BMVC*, pp. 1–11, Sep. 2009.

[34] J. Liu, J. Luo, and M. Shah, “Recognizing realistic actions from videos in the wild,” in *Proc. IEEE CVPR*, pp. 1996–2003, Jun. 2009.

[35] Q. V. Le, W. Y.Zhou, S. Y. Yeung, and A. Y. Ng, “Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis,” in *Proc. IEEE CVPR*, pp. 3361–3368, Jun. 2011.

[36] A. H. Shabani, D. A. Clausi, and J. S. Zelek, “Improved spatio-temporal salient feature detection for action recognition,” in *Proc. BMVC*, pp. 1–12, Sep. 2011.

[37] L. Yeffet and L. Wolf, “Local trinary patterns for human action recognition,” in *Proc. IEEE ICCV*, pp. 492–497, Sep. 2009.

[38] S. Wu, O. Oreifej, and M. Shah, “Action recognition in videos acquired by a moving camera using motion decomposition of Lagrangian particle trajectories,” in *Proc. IEEE ICCV*, pp. 1419–1426, Nov. 2011.

[39] M. Bregonzio, J. Li, and S. Gong, “Discriminative topics modelling for action feature selection and recognition,” in *Proc. BMVC*, pp. 1–11, Oct. 2010.

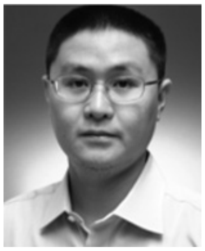
[40] B. Chakraborty, B. M. Holte, T. B. Moeslund, and J. Gonzalez, “Selective spatio-temporal interet points,” *Comput. Vision Image Understanding*, vol. 116, no. 3, pp. 396–410, Mar. 2012.

[41] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May–Jun. 2001.

[42] I. Laptev, “On space-time interest points,” *Int. J. Comput. Vision*, vol. 64, no. 2–3, pp. 107–123, Sep. 2005.

Xinxiao Wu received the B.S. degree from the Nanjing University of Information Science and Technology, Nanjing, China, in 2005, and the Ph.D. degree from the Beijing Institute of Technology, Beijing, China, in 2010.

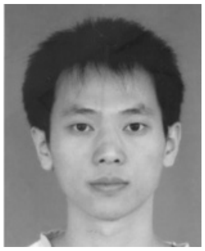
She was a Post-Doctoral Research Fellow with Nanyang Technological University, Singapore, for one year. She is currently a Lecturer with the Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology. Her current research interests include computer vision, machine learning, and video content analysis.



Dong Xu (M’07) received the B.E. and Ph.D. degrees from the University of Science and Technology of China, Hefei, China, in 2001 and 2005, respectively.

While pursuing the Ph.D. degree, he was with Microsoft Research Asia, Beijing, China, and the Chinese University of Hong Kong, Shatin, Hong Kong, for more than two years. He was a Post-Doctoral Research Scientist with Columbia University, New York, NY, USA, for one year. In May 2007, he joined Nanyang Technological University, Singapore, where he is currently an Associate Professor with School of Computer Engineering. His current research interests include computer vision, statistical learning, and multimedia content analysis.

Dr. Xu was the co-author of a paper that won the Best Student Paper Award in the prestigious IEEE International Conference on Computer Vision and Pattern Recognition in 2010.



Lixin Duan received the B.E. degree from the University of Science and Technology of China, Hefei, China, in 2008, and the Ph.D. degree from Nanyang Technological University, Singapore, in 2012.

He is currently a Researcher with SAP Next Business and Technology, Singapore.

Dr. Duan was a recipient of the Microsoft Research Asia Fellowship in 2009 and the Best Student Paper Award at the IEEE Conference on Computer Vision and Pattern Recognition in 2010.

Jiebo Luo (S'93–M'96–SM'99–F'09) received the B.S. degree from the University of Science and Technology of China Hefei, China, in 1989, and the Ph.D. degree from the University of Rochester, Rochester, NY, USA, in 1995.

He was a Senior Principal Scientist with the Kodak Research Laboratories, Rochester, before joining the Computer Science Department at the University of Rochester in 2011. He has authored over 200 technical papers and holds over 70 U.S. patents. His current research interests include image processing, machine learning, computer vision, social multimedia data mining, biomedical informatics, and ubiquitous computing.

Dr. Luo is also a Fellow of the SPIE and IAPR.

Yunde Jia (M'11) received the M.S. and Ph.D. degrees in mechatronics from the Beijing Institute of Technology (BIT), Beijing, China, in 1986 and 2000, respectively.

He is currently a Professor of computer science with BIT, and serves as the Director of the Beijing Laboratory of Intelligent Information Technology, School of Computer Science. He has previously served as the Executive Dean of the School of Computer Science, BIT, from 2005 to 2008. He was a Visiting Scientist at Carnegie Mellon University, Pittsburgh, PA, USA, from 1995 to 1997, and a Visiting Fellow at the Australian National University, Acton, Australia, in 2011. His current research interests include computer vision, media computing, and intelligent systems.