Action Recognition using Context and Appearance Distribution Features

Xinxiao Wu Dong Xu Lixin Duan
School of Computer Engineering
Nanyang Technological University
Singapore
{xinxiaowu,DongXu,S080003}@ntu.edu.sg

Jiebo Luo
Kodak Research Laboratories
Eastman Kodak Company
Rochester,USA
Jiebo.Luo@kodak.com

Abstract

We first propose a new spatio-temporal context distribution feature of interest points for human action recognition. Each action video is expressed as a set of relative XYT coordinates between pairwise interest points in a local region. We learn a global GMM (referred to as Universal Background Model, UBM) using the relative coordinate features from all the training videos, and then 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 multi-scale local regions. To describe the appearance information of an action video, we also propose to use GMM to characterize the distribution of local appearance features from the cuboids centered around the interest points. Accordingly, an action video can be represented by two types of distribution features: 1) multiple GMM distributions of spatio-temporal context; 2) GMM distribution of local video appearance. To effectively fuse these two types of heterogeneous and complementary distribution features, we additionally propose a new learning algorithm, called Multiple Kernel Learning with Augmented Features (AFMKL), to learn an adapted classifier based on multiple kernels and the pre-learned classifiers of other action classes. Extensive experiments on KTH, multi-view IXMAS and complex UCF sports datasets demonstrate that our method generally achieves higher recognition accuracy than other state-of-the-art methods.

1. 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. How to extract discriminative and robust image features to describe actions and design new effective learning methods to fuse different types of features have become two important issues in action recognition.

Human action recognition algorithms can be roughly categorized into model-based methods and appearance-based approaches. Model-based methods [5, 29] 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 [7, 30] to describe an action. However, with these methods, highly detailed silhouettes need to be extracted, which may be very difficult in a realistic video. Recently, approaches [3, 21, 16] 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 3D 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 [3, 16, 21] model an action as a bag of independent and orderless visual words without considering the spatio-temporal contextual information of interest points.

In Section 3, we first propose a new visual feature by using multiple 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., sub-volume) 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 a set of context features extracted from all pairs of interest points over all the local regions in a video volume. Gaussian Mixture Model (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 work, multiple GMMs are exploited to cope with different levels of spatio-temporal contexts of interest points from different space-time scales. GMM is also used to characterize the appearance distribution of local cuboids that are centered around the interest points.

Multi-scale spatio-temporal context distribution feature and appearance distribution feature characterize the properties of "where" and "what" of interest points, respectively. In Section 4, we propose a new learning method called Multiple Kernel Learning with Augmented Features (AFMKL) to effectively integrate two types of complementary features. Specifically, we propose to learn an adapted classifier based on multiple kernels constructed from different types of features and the pre-learned classifiers from other action classes. AFMKL is motivated by the observation that some actions share similar motion patterns (e.g., the actions of "walking", "jogging" and "running" may share some similar motions of hands and legs). It is beneficial to learn an adapted classifier for "jogging" by leveraging the pre-learned classifiers from "walking" and "running". It is interesting to observe that the dual of our new objective function is similar to that of Generalized Multiple Kernel Learning (GMKL) [23] except that the kernel matrix is computed based on the Augmented Features, which combines the original context/appearance distribution feature and the decision values from the pre-learned SVM classifiers of all the classes.

In Section 5, we conduct comprehensive experiments on the KTH, multi-view IXMAS and UCF sports datasets, and the results demonstrate the effectiveness of our method. The main contributions of this work include: 1) we propose a new multi-scale spatio-temporal context distribution feature; 2) we propose a new learning method AFMKL to effectively fuse the context distribution feature and the appearance distribution feature; 3) our method generally outperforms the existing methods on three benchmark datasets, demonstrating promising results for realistic action recognition.

2. Related Work

Recently, researchers have exploited the spatial and temporal context as another type of information for describing interest points. Kovashka and Grauman [11] exploited multiple "bag-of-words" models to represent the hierarchy of space-time configurations at different scales. Savarese *et al*. [20] 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 *et al*. [19] proposed a so-called "featuretype×featuretype×relationship" histogram to capture both appearance and relationship information between pairwise visual words. All these methods first utilized the vector quantization process to generate a codebook and then adopted the "bag-of-words" representation. The quantization error may be propagated to the spatio-temporal context features and may degrade the final recognition performance. In our work, the spatio-temporal context feature is directly extracted from the detected interest points rather than visual words, so our method dose not suffer from the vector quantization error. Moreover, the 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 the video-specific GMM adapted from global GMM provides additional information to distinguish the videos of different classes.

Sun *et al*. [22] presented a hierarchical structure to model the spatio-temporal context information of SIFT points, and their model consists of point-level context, intra-trajectory context and inter-trajectory context. Bregonzio *et al*. [1] created the 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*. [31] extracted the motion words from the motion images and utilized the relative locations between the motion words and a reference point in a local region to establish the spatio-temporal context information. These methods mentioned above are based on various pre-processing steps, such as feature tracking, human body detection and foreground segmentation. As will be shown without requiring any pre-processing step, our method can still achieve promising performance even in complex environments with changing lighting and moving cameras.

Vega *et al*. [24] and Nayak *et al*. [15] proposed to use the distribution of pairwise relationships between the edge primitivies to capture the shape of action in 2D image. In contrast to [24] [15], the relationship between the interest points in this paper is defined in 3D video volume to capture both motion and shape of action. Moreover, they described the rational distribution just within the whole image, while we used multiple GMMs to model the context distribution of interest points at multiple space-time scales.

3. Spatio-Temporal Context Distribution Feature and Appearance Distribution Feature

3.1. Detection of interest points

We use the interest point detector proposed by Dollar *et al*. [3] which respectively employs 2D Gaussian filter in the spatial direction and 1D Gabor filter in the temporal direction. The two separate filters can produce high response at

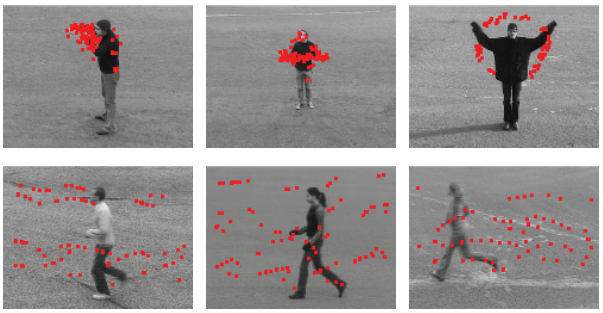

Figure 1. Samples of detected interest points on the KTH dataset.

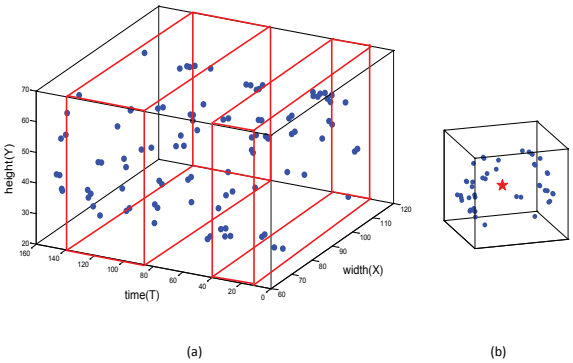

Figure 2. The multi-scale spatio-temporal context extraction from interest points of one "handwaving" video in a 3D video volume (a) and in a local region (b).

points with significant spatio-temporal intensity variations. The response function has the form:

$$(I(x,y,t)*g(x,y)*h_{ev}(t))^2+(I(x,y,t)*g(x,y)*h_{od}(t))^2,$$

where $g(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}}$ is the 2D Gaussian smoothing kernel along the spatial dimension, and $h_{ev}(t) = -cos(2\pi t\omega)e^{-t^2/\tau^2}$ and $h_{od}(t) = -sin(2\pi t\omega)e^{-t^2/\tau^2}$ are a pair of 1D Gabor filters along the temporal dimension. With the constraints $\omega = 4/\tau$, σ and τ are two parameters of the spatial scale and temporal scale, respectively. Figure 1 shows some examples of detected interest points (depicted by red dots) of human actions. It is evident that most detected interest points are near the body parts that have major contribution to the action classification.

3.2. Multi-scale 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 by:

$$[X_c\ Y_c\ T_c] = \arg\min_{[X_c,Y_c,T_c]} \sum_{i=1}^{R} \|[X_i\ Y_i\ T_i]^T - [X_c\ Y_c\ T_c]^T\|^2,$$

where $[X_c\ Y_c\ T_c]^T$ and $[X_i\ Y_i\ T_i]^T$ respectively represent the coordinates of the center and the i-th 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., $s_i = [X_i - X_c\ Y_i - Y_c\ T_i - T_c]^T$, $i = 1, 2, ..., R$. As shown in Figure 2(b), the red pentacle 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 multi-scale local regions across multiple space-time scales to generate multiple sets of local context features (i.e., XYT relative coordinates). Each set represents the spatio-temporal context at one space-time scale. Figure 2(a) illustrates the detected interest points of one video from the "handwaving" action, where the blue dots are the interest points in 3D 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 $N = RT$.

3.3. Multi-scale spatio-temporal context distribution feature

For each action video, a video-specific Gaussian Mixture Model (GMM) is employed to characterize the distribution of the set of spatio-temporal context features at one space-time scale. 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 Universal Background Modeling, 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 k-th Gaussian component, respectively. Note we have the constraint $\sum_{k=1}^K m_k = 1$. As suggested in [32], the covariance matrix Σ_k is set to be a diagonal matrix for computational efficiency. We adopt the well-known Expectation-Maximization (EM) 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 a Maximum A Posterior (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 i-th 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 k-th 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 k-th GMM component. Then, the k-th component of the video-specific GMM of any video V can be adapted as follows:

$$\begin{aligned} \bar{\mu}_k &= \frac{\sum_{i=1}^N \eta(k|s_i) s_i}{\eta_k} \\ \hat{\mu}_k &= (1-\xi_k)\mu_k + \xi_k \bar{\mu}_k \\ \hat{m}_k &= \frac{\zeta_k}{N}, \end{aligned}$$

where $\bar{\mu}_k$ is the expected mean of the k-th component based on the training samples $s_i|_{i=1}^N$ and $\hat{\mu}_k$ is the adapted mean of the k-th component. The weighting coefficient $\xi_k = \frac{\zeta_k}{\zeta_k + r}$ is introduced to improve the estimation accuracy of the mean vector. If a Gaussian component has a high soft count η_k, then ξ_k approaches 1 and the adapted mean is mainly decided by the statistics of training samples; otherwise, the adapted mean is mainly determined by the global model. Following [32], 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 k-th 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.

3.4. Local video appearance distribution feature

After detecting the interest points, we also extract the appearance information from the cuboids around the interest points. For simplicity, we flatten each normalized cuboid and extract the gray-level pixel values from each normalized cuboid. Principle Component Analysis (PCA) is used to reduce the dimension of appearance feature vector by preserving 98% energy. Similar to the spatio-temporal context distribution feature, we also employ GMM to describe the distribution of local appearance for each action video by using the two-step approach discussed in Section 3.3.

Discussion: In [32], the GMM is adopted to represent the distribution of SIFT features in a video for event recognition. While SIFT is a type of static feature extracted from 2D image patches, our cuboid feature captures both static and motion information from 3D video sub-volume which is more effective for describing actions.

4. Multiple Kernel Learning with Augmented Features (AFMKL) for Action Recognition

To fuse the spatio-temporal context distribution and local appearance distribution features for action recognition, we propose a new learning method called Multiple Kernel Learning with Augmented Features (AFMKL) to learn a robust classifier by using multiple base kernels and a set of pre-learned SVM classifiers from other action classes. The introduction of the pre-learned classifiers from other classes is based on the observation that some actions may share common motion patterns. For example, the actions of "walking" , "jogging" and "running" may share some typical motions of hands and legs, therefore it is beneficial to learn an adapted classifier for "jogging" by leveraging the pre-learned classifiers of "walking" and "running".

Specially, in AFMKL, we use M base kernel functions $k_m(x_i, x_j) = \varphi_m(x_i)^T \varphi_m(x_j)$, where $m = 1, 2, ..., M$, and x_i, x_j can be the context or appearance distribution feature extracted from video V_i and V_j, respectively. In this work, we use linear kernels, i.e., $k_m(x_i, x_j) = x_i^T x_j$, and we have $M = H + 1$ where H is the number of space-time scales in the context distribution features. Multiple kernel functions are linearly combined to determine an optimal kernel function k, i.e., $k = \sum_{m=1}^M d_m k_m$, where d_m's

are the linear combination coefficients, $\sum_{m=1}^{M} d_m = 1$ and $d_m \geq 0$. For each class, we have one pre-learned SVM classifier using the appearance feature and one pre-learned SVM classifier by fusing H context features from different space-time scales in an early fusion fashion. Suppose we have C classes, then we have $L = 2C$ pre-learned classifiers $\{f_l(x)|_{l=1}^{L}\}$ from all the classes. Given the input feature x of the video, the final decision function is defined as follow:

$$f(x) = \sum_{l=1}^{L} \beta_l f_l(x) + \sum_{m=1}^{M} d_m \mathbf{w}_m^T \varphi_m(x) + b, \qquad (1)$$

where β_l is the weight of the l-th pre-learned classifier, $\mathbf{w}_m$ and b are the parameters of the standard SVM. Note that $\sum_{m=1}^{M} d_m \mathbf{w}_m^T \varphi_m(x) + b$ is the decision function in Multiple Kernel Learning (MKL).

Given the training videos $\{x_i|_{i=1}^{I}\}$ with the corresponding class labels $\{y_i|_{i=1}^{I}\}$ where $y_i \in \{-1, +1\}$, we formulate AFMKL as follows by adding a quadratic regularization term on d_m's:

$$\min_{\mathbf{d}} \ \frac{1}{2}\|\mathbf{d}\|^2 + J(\mathbf{d}), \quad s.t. \ \sum_{m=1}^{M} d_m = 1, \ d_m \geq 0, \qquad (2)$$

where

$$J(\mathbf{d}) = \left\{ \begin{array}{l} \min\limits_{\mathbf{w}_m, \boldsymbol{\beta}, b, \xi_i} \ \frac{1}{2}\left(\sum\limits_{m=1}^{M} d_m \|\mathbf{w}_m\|^2 + \lambda \|\boldsymbol{\beta}\|^2 \right) + C \sum\limits_{i=1}^{I} \xi_i, \\ s.t. \quad y_i f(x_i) \geq 1 - \xi_i, \ \xi_i \geq 0, \ i = 1, \ldots, I, \end{array} \right.$$

$\mathbf{d} = [d_1, d_2, \ldots, d_M]^T$ is the vector of linear combination coefficients and $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_L]^T$ is the weighting vector. Note that in $J(\mathbf{d})$ we penalize the complexity of the weighting vector $\boldsymbol{\beta}$ to control the complexity of the pre-learned classifiers. By replacing $\mathbf{w}_m = \frac{\tilde{\mathbf{w}}_m}{d_m}$ and according to [17], the above optimization problem in (2) is jointly convex with respect to $\mathbf{d}$, $\tilde{\mathbf{w}}_m$, $\boldsymbol{\beta}$, b and ξ_i. Thus, the global optimum of the objective in (2) can be reached. Similarly as in [23], by replacing the primal form of the optimization problem in $J(\mathbf{d})$ with its dual form, $J(\mathbf{d})$ can be rewritten as:

$$J(\mathbf{d}) = \left\{ \begin{array}{l} \max\limits_{\boldsymbol{\alpha}} \ \sum\limits_{i=1}^{I} \alpha_i - \frac{1}{2} \sum\limits_{i,j=1}^{I} \alpha_i \alpha_j y_i y_j \sum\limits_{m=1}^{M} d_m \tilde{k}_m(x_i, x_j), \\ s.t. \ \sum\limits_{i=1}^{I} \alpha_i y_i = 0, \ 0 \leq \alpha_i \leq C, \ i = 1, \ldots, I, \end{array} \right.$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \ldots, \alpha_I]^T$ is the vector of the dual variables α_i's and $\tilde{k}_m(x_i, x_j) = \varphi(x_i)^T \varphi(x_j) + \frac{1}{\lambda} \sum_{l=1}^{L} f_l(x_i) f_l(x_j)$. It is interesting to observe that the dual problem of (2) is similar to that of Generalized MKL [23] except that the kernel function is replaced by $\sum_{m=1}^{M} d_m \tilde{k}_m$ in our work. Since we use linear kernel in this work (i.e., $\varphi(x_i) = x_i$), then $\tilde{k}_m(x_i, x_j)$ can be computed using the augmented features $[x_i^T, \frac{1}{\sqrt{\lambda}} f_1(x_i), ..., \frac{1}{\sqrt{\lambda}} f_L(x_i)]^T$ and $[x_j^T, \frac{1}{\sqrt{\lambda}} f_1(x_j), ..., \frac{1}{\sqrt{\lambda}} f_L(x_j)]^T$ which combine the original context/appearance feature and the decision values from the pre-learned SVM classifiers of all the classes. We therefore name our method as Multiple Kernel Learning with Augmented Features (AFMKL).

Following [23], we iteratively update the linear combination coefficient $\mathbf{d}$ and the dual variable $\boldsymbol{\alpha}$ to solve (2). In [23], the first-order gradient descent method is used for updating $\mathbf{d}$ in (2). In contrast, we employ the second-order gradient descent method because of its faster convergence. After obtaining the optimal $\mathbf{d}$ and $\boldsymbol{\alpha}$, we rewrite the decision function in (1) as follows:

$$f(x) = \sum_{i=1}^{I} \alpha_i y_i \left(\sum_{m=1}^{M} d_m k_m(x_i, x) + \frac{1}{\lambda} \sum_{l=1}^{L} f_l(x_i) f_l(x) \right) + b.$$

Discussion: The most related method is Adaptive Multiple Kernel Learning (A-MKL) [4]. A-MKL is proposed for cross-domain learning problems and it learns a target classifier by leveraging the pre-learned classifiers from *the same class*, which are trained based on the training data from *two domains* (i.e., auxiliary domain and target domain). In contrast, our method AFMKL is specifically proposed for action recognition and all the samples are assumed to be from *the same domain*. The pre-learned classifiers used in AFMKL are from *other action classes*. Our method AFMKL is also different from SVM with Augmented Features (AFSVM) proposed in [2]. Although AFSVM also makes use of the pre-learned classifiers from other classes, only one kernel is considered in AFSVM. In contrast, our proposed method AFMKL is a Multiple Kernel Learning (MKL) technique that can learn the optimal kernel by linearly combining multiple kernels constructed from different types of features.

5. Experiments

5.1. Human action datasets

The KTH human dataset [21] is a commonly used action dataset. It contains six human action classes: walking, jogging, running, boxing, handwaving and handclapping. These actions are performed by 25 subjects in four different scenarios: outdoors (s1), outdoors with scale variation (s2), outdoors with different clothes (s3) and indoors with lighting variation (s4). There are totally 599 video clips with the image size of 160×120 pixels. We adopt the leave-one-subject-out cross validation setting in which videos of 24 subjects are used as training data and the videos of the remaining one subject are used for testing.

The IXMAS dataset [26] consists of 12 complete action classes with each action executed three times by 12 subjects and recorded by five cameras with the frame size of 390×291 pixels. These actions are: check watch, cross arms, scratch head, sit down, get up, turn around, walk, wave, punch, kick, point and pick up. The body position and orientation are freely decided by different subjects. As in the KTH action dataset, we also use the same leave-one-subject-out cross validation setting.

The UCF sports dataset [18] contains ten different types of sports actions: swinging, diving, kicking, weight-lifting, horse-riding, running, skateboarding, swinging at the high bar (HSwing), golf swinging (GSwing) and walking. The dataset consists of 149 real videos with a large intra-class variability. 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 [25]. 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 [25], the flipped version of the test video sequence is not included in the training set.

5.2. Experimental setting

For interest points detection, the spatial and temporal scale parameters σ and τ are empirically set by $\sigma = 2.5$ and $\tau = 2$, respectively. The size of the cuboid is empirically fixed as $7 \times 7 \times 5$. For the KTH and IXMAS datasets, 200 interest points are extracted from each video, because they contain enough information to distinguish the actions in relatively homogeneous and static background. About 1000 interest points detected from each video are used for the more complex UCF sports dataset. The number of Gaussian components in GMM (i.e., K) is empirically set to 300, 400 and 300 for the KTH, IXMAS and UCF datasets, respectively.

For our method, we report the results using SVM with multi-scale spatio-temporal (ST) context distribution feature only, SVM with local appearance distribution feature only, the combinations of these two types of features via Generalized MKL (GMKL) [23] and AFMKL. For SVM with ST context distribution feature, we adopt the early fashion method to integrate the multiple features from H space-time scales. In both GMKL and AFMKL, we use $M = H + 1$ linear base kernels respectively constructed for the sptaio-temporal context distribution features from H space-time scales and the appearance distribution feature. To cope with multi-class classification task using SVM, we adopt the default setting in LIBSVM.

Method	Accuracy(%)
Dollar *et al*. [3]	81.2
Savarese *et al*. [20]	86.8
Niebles *et al*. [16]	81.5
Liu and Shah [14]	94.3
Bregonizo *et al*. [1]	93.2
Ryoo and Aggarwal [19]	91.1
Schuldt *et al*. [21]	71.7
JHuang *et al*. [8]	91.7
Klaser *et al*. [10]	84.3
Zhang *et al*. [31]	91.3
Laptev *et al*. [12]	91.8
Liu *et al*. [13]	93.8
Gilbert *et al*. [6]	**94.5**
Kovashka and Grauman [11]	**94.5**
Our method	**94.5**

Table 1. Recognition accuracies(%) of different methods on the KTH dataset.

Method	S1	S2	S3	S4	Ave
ST Context	92.7	76.7	86.0	88.7	86.0
Appearance	90.7	89.3	87.3	90.7	89.5
GMKL [23]	96.0	86.0	90.7	94.0	91.7
AFMKL	**96.7**	**91.3**	**93.3**	**96.7**	**94.5**

Table 2. Recognition accuracies (%) using different features on four scenarios of the KTH dataset. The last column is the average accuracy of all scenarios.

Figure 3. Confusion table of our AFMKL on the KTH dataset.

Camera	1	2	3	4	5
ST Context	61.1	57.9	53.2	51.6	42.6
Appearance	72.2	72.2	70.4	65.7	67.6
GMKL [23]	76.4	74.5	73.6	71.8	60.4
AFMKL	**81.9**	**80.1**	**77.1**	**77.6**	**73.4**
Liu and Shah [14]	76.7	73.3	72.1	73.1	-
Yan *et al*. [27]	72.0	53.0	68.1	63.0	-
Weinland *et al*. [26]	65.4	70.0	54.4	66.0	33.6
Junejo *et al*. [9]	76.4	77.6	73.6	68.8	66.1

Table 3. Classification accuracies (%) of different methods for single view action recognition on the IXMAS dataset.

5.3. Experimental results

Table 1 compares our method with the existing methods on the KTH dataset. Among the results [3, 20, 16, 14, 1, 19]

Cameras	1,2,3,4,5	1,2,3,4	1,2,3,5	1,2,3	1,3,5	1,3	2,4	3,5
ST Context	73.8	72.2	73.2	73.2	67.8	67.4	66.7	57.6
Appearance	81.9	77.6	83.6	77.8	82.4	74.3	72.7	79.9
GMKL [23]	81.3	80.8	81.5	81.3	77.6	76.2	77.1	72.7
AFMKL	**88.2**	**88.2**	**89.4**	**87.7**	**88.4**	**86.6**	**82.4**	**83.8**
Liu and Shah [14]	82.8	-	-	-	-	-	-	-
Yan *et al*. [27]	-	78.0	-	60.0	-	71.0	71.0	-
Weinland *et al*. [26]	-	81.3	75.9	-	70.2	-	81.3	61.6

Table 4. Classification accuracies (%) of different methods for multi-view action recognition on the IXMAS dataset.

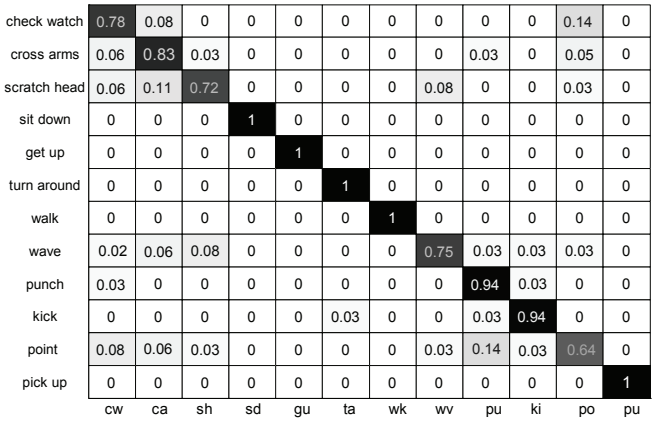

Figure 4. Confusion table of our AFMKL in the multi-view setting using all five views on the IXMAS dataset.

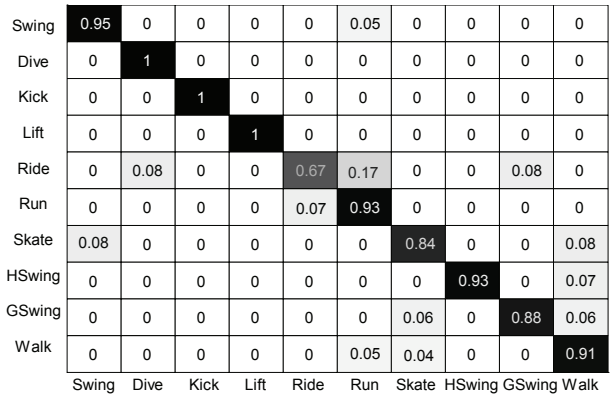

Figure 5. Confusion table of our AFMKL on the UCF sports dataset.

Method	Accuracy (%)
ST Context	71.1
Appearance	76.5
GMKL [23]	85.2
AFMKL	**91.3**
Kovashka and Grauman [11]	87.3
Wang *et al*. [25]	85.6
Rodriguez *et al*. [18]	69.2
Yeffet and Wolf [28]	79.3

Table 5. Recognition accuracies (%) of different methods on the UCF sports dataset.

using the interest points detected by Dollar's method [3] and the leave-one-subject-out cross validation setting, our method achieves the highest recognition accuracy of 94.5%, which is also the same as the best results from the recent works [21, 8, 10, 31, 12, 13, 6, 11]. Table 2 reports the recognition accuracies on four different scenarios. Figure 3 shows the confusion table of recognition results on the KTH dataset. It is interesting to note that the leg-related actions ("jogging", "running", "walking") are more confused with each other. Especially "running" is easy to be misclassified as "jogging". The possible explanation is that "jogging" and "running" have similar spatio-temporal context and appearance information.

Our approach is also tested for single-view and multi-view action recognition on the IXMAS dataset. Table 3 and Table 4 respectively report the classification results of different methods for single-view and multi-view action recognition. Among the methods [14, 26, 27, 9] that are tested on the IXMAS dataset, the method [14] is the most related one because the spatio-temporal interest points are also used in [14]. Other methods need to use more information of 3D human poses via camera calibration, background subtraction and 3D pose construction which are not employed in this work. The confusion table of recognition results in the multi-view setting using all five views is shown in Figure 4. For some actions, such as "sit down", "get up", "turn around", "walk" and "pick up", our method achieves very high recognition accuracies. Our method also achieves reasonable performances for some challenging actions (e.g., "point", "scratch head" and "wave") that have small and ambiguous motions.

Evaluation results on the UCF sports videos are presented in Table 5. Our method achieves 91.3% recognition accuracy which is better than 85.6% and 87.3% respectively reported in recent work [25] and [11] using the same setting. The confusion matrix of our AFMKL is depicted in Figure 5. While this dataset is difficult with large viewpoint and appearance variability as well as camera motion, our result is still encouraging.

From Tables 2, 3, 4 and 5, it is important to note that although the spatio-temporal context distribution feature is

generally not as effective as the appearance distribution feature, the combination of two types of complementary features via GMKL [23] significantly outperforms the methods using single feature in most cases. Moreover, AFMKL achieves the best results in all the cases, which demonstrates the effectiveness of using the pre-learned classifiers from other action classes and multiple kernel learning to improve the recognition performance.

6. Conclusions

In order to describe the "where" property of interest points, we have proposed a new visual feature by using multiple Gaussian Mixture Models (GMMs) to represent the distributions of local spatio-temporal context between interest points at different space-time scales. The local appearance distribution in each video is also modeled using one GMM in order to capture the "what" property of interest points. To fuse both spatio-temporal context distribution and local appearance distribution features, we additionally propose a new learning algorithm called Multiple Kernel Learning with Augmented Features (AFMKL) to learn an adapted classifier by leveraging the existing SVM classifiers of other action classes. Extensive experiments on KTH, IXMAS and UCF datasets have demonstrated that our method generally outperforms the state-of-the-art algorithms for action recognition. In the future, we plan to investigate how to automatically determine the optimal parameters in our method.

Acknowledgements This work is funded by Singapore A*STAR SERC Grant (082 101 0018).

References

[1] M. Bregonzio, S. Gong, and T. Xiang. Recognising action as clouds of space-time interest points. In *CVPR*, 2009.

[2] L. Chen, D. Xu, I. W.-H. Tsang, and J. Luo. Tag-based web photo retrieval improved by batch mode re-tagging. In *CVPR*, 2010.

[3] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS PETS*, 2005.

[4] L. Duan, D. Xu, I. W.Tsang, and J. Luo. Visual event recognition in videos by learning from web data. In *CVPR*, 2010.

[5] C. Fanti, L. Zelnik-manor, and P. Perona. Hybrid models for human motion recognition. In *ICCV*, 2005.

[6] A. Gilbert, J. Illingworth, and R. Bowden. Fast realistic multi-action recognition using mined dense spatio-temporal features. In *ICCV*, 2009.

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

[8] H. Jhuang, T. Serre, L. Wolf, and T. Poggio. A biologically inspired system for action recognition. In *ICCV*, 2007.

[9] I. N. Junejo, E. Dexter, I. Laptev, and P. Prez. Cross-view action recognition from temporal self-similarities. In *ECCV*, 2008.

[10] A. Klaser, M. Marszalek, and C. Schmid. A spatio-temporal descriptor based on 3d-gradients. In *BMVC*, 2008.

[11] A. Kovashka and K. Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *CVPR*, 2010.

[12] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, I. Rennes, I. I. Grenoble, and L. Ljk. Learning realistic human actions from movies. In *CVPR*, 2008.

[13] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos. In *CVPR*, 2009.

[14] J. Liu and M. Shah. Learning human actions via information maximization. In *CVPR*, 2008.

[15] S. Nayak, S. Sarkar, and B. L. Loeding. Distribution-based dimensionality reduction applied to articulated motion recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(5):795–810, 2009.

[16] J. C. Niebles, H. Wang, and L. Fei-fei. Unsupervised learning of human action categories using spatial-temporal words. In *IJCV*, volume 79, pages 299–318, 2008.

[17] A. Rakotomamonjy, F. R. Bach, and Y. Grandvalet. SimpleMKL. *JMLR*, 9:2491–2521, 2008.

[18] M. D. Rodriguez, J. Ahmed, and M. Shah. Action mach a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008.

[19] M. Ryoo and J. Aggarwal. Spatio-temporal relationship match: video structure comparison for recognition of complex human activities. In *ICCV*, 2009.

[20] S. Savarese, A. Delpozo, J. C. Niebles, and L. Fei-fei. Spatial-temporal correlatons for unsupervised action classification. In *WMVC*, 2008.

[21] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004.

[22] J. Sun, X. Wu, S. Yan, L. F. Cheong, T.-S. Chua, and J. Li. Hierarchical spatio-temporal context modeling for action recognition. In *CVPR*, 2009.

[23] M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *ICML*, 2009.

[24] 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.*, 25(10):1323–1328, 2003.

[25] H. Wang, M. M. Ullah, A. Klser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009.

[26] D. Weinland, E. Boyer, and R. Ronfard. Action recognition from arbitrary views using 3d exemplars. In *ICCV*, 2007.

[27] P. Yan, S. M. Khan, and M. Shah. Learning 4d action feature models for arbitrary view action recognition. In *CVPR*, 2008.

[28] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009.

[29] A. Yilmaz. Recognizing human actions in videos acquired by uncalibrated moving cameras. In *ICCV*, 2005.

[30] A. Yilmaz and M. Shah. Actions sketch: a novel action representation. In *CVPR*, 2005.

[31] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia. Motion context: a new representation for human action recognition. In *ECCV*, 2008.

[32] X. Zhou, X. Zhuang, S. Yan, S.-F. Chang, M. Hasegawa-Johnson, and T. S. Huang. Sift-bag kernel for video event analysis. In *ACM Multimedia*, 2008.