Sensors **2013**, *13*, 14398-14416; doi:10.3390/s131114398

OPEN ACCESS

sensors

ISSN 1424-8220

www.mdpi.com/journal/sensors

Article

A Generalized Pyramid Matching Kernel for Human Action Recognition in Realistic Videos

Jun Zhu [1], Quan Zhou [2], Weijia Zou [1], Rui Zhang [1] and Wenjun Zhang [1,*]

[1] Institute of Image Communication and Information Processing, Shanghai Jiao Tong University, Shanghai 200240, China; E-Mails: zhujun.sjtu@gmail.com (J.Z.); zouweijia@sjtu.edu.cn (W.Z.); zhang_rui@sjtu.edu.cn (R.Z.)

[2] College of Telecommunications and Information Engineering, Nanjing University of Posts and Telecommunications, Nanjing 210003, China; E-Mail: quan.zhou@njupt.edu.cn

* Author to whom correspondence should be addressed; E-Mail: zhangwenjun@sjtu.edu.cn; Tel.: +86-21-3420-6596; Fax: +86-21-3420-4155.

Received: 6 September 2013; in revised form: 30 September 2013; / Accepted: 5 October 2013 / Published: 24 October 2013

Abstract: Human action recognition is an increasingly important research topic in the fields of video sensing, analysis and understanding. Caused by unconstrained sensing conditions, there exist large intra-class variations and inter-class ambiguities in realistic videos, which hinder the improvement of recognition performance for recent vision-based action recognition systems. In this paper, we propose a generalized pyramid matching kernel (GPMK) for recognizing human actions in realistic videos, based on a multi-channel "bag of words" representation constructed from local spatial-temporal features of video clips. As an extension to the spatial-temporal pyramid matching (STPM) kernel, the GPMK leverages heterogeneous visual cues in multiple feature descriptor types and spatial-temporal grid granularity levels, to build a valid similarity metric between two video clips for kernel-based classification. Instead of the predefined and fixed weights used in STPM, we present a simple, yet effective, method to compute adaptive channel weights of GPMK based on the kernel target alignment from training data. It incorporates prior knowledge and the data-driven information of different channels in a principled way. The experimental results on three challenging video datasets (*i.e.*, Hollywood2, Youtube and HMDB51) validate the superiority of our GPMK *w.r.t.* the traditional STPM kernel for realistic human action recognition and outperform the state-of-the-art results in the literature.

Keywords: video analysis; human action recognition; pyramid matching kernel; kernel-based classification method

1. Introduction

Recognition of human actions, e.g., running, fighting and shooting balls, is an increasingly important research topic in the fields of video sensing, analysis and understanding [1–3]. It applies to a wide range of computer vision applications, such as video surveillance [4,5], human-computer interaction [6,7], sports video analysis [8,9] and content-based video retrieval [10,11]. Although promising progress has been achieved for human action recognition in constrained scenarios [12,13], recognition accuracy remains unsatisfactory for realistic videos (e.g., TV, movies and Internet videos) [14–16]. This is mainly because they are taken under unconstrained sensing conditions and, thus, suffer from a great number of visual challenges (e.g., object pose, background clutter, camera motion, viewpoint and illumination variations), which result in large intra-class variations and inter-class ambiguities that hinder the improvement of recognition performance in recent vision-based action recognition systems.

To overcome these challenges, massive research efforts have been dedicated to vision-based systems on realistic human action recognition over last few years [9,14–20]. Related work in human action recognition literature can be generally divided into two categories: (1) The first category relies on the technologies of detecting and analyzing human body movement (e.g., kinematic tracking [21], human body pose estimation [22], space-time shape templates [13], *etc*.) in video sequences and, then, performs action recognition on that basis; (2) As an extension of a classic framework in the image classification field [23–27], the second category aims at directly building a holistic feature representation of the video clip for human action recognition, based on *local spatial-temporal features* [28] and the "*bag of words*" (BoW) model [23]. Specifically, we can first extract localized feature points from video clips by using spatial-temporal interest point (STIP) detectors [17,29] or dense sampling methods [20,28] and, then, capture appearance, shape and motion information in the neighborhoods of detected points by feature descriptors [17,20,30–32]. After that, the BoW model calculates the visual word frequency statistics on quantized local feature points, to form a histogram-wise video representation for classification. Compared with the first category of action recognition approaches, the latter one becomes increasingly dominant for realistic human action recognition because of its many advantages: the detected STIPs have a certain degree of invariance to illumination, scale and viewpoint changes; the BoW representation is robust to relatively large intra-class variations of human action patterns in realistic videos; it is easy to be implemented and does not require complicated motion capture or human body analysis technologies.

However, traditional BoW representation fails to capture useful distribution information in the spatial-temporal domain for detected orderless STIPs, which is crucial for the recognition of complex action classes in realistic videos. The success of *spatial pyramid matching* (SPM) [24] in the static image recognition domain [25–27] motivates researchers to incorporate the spatial-temporal layout information of local features for dynamic videos [10,17,19,20,33]. Choi *et al*. [10] extended the

SPM technology into the scenario of 3D video clips and proposed a *spatial-temporal pyramid matching* (STPM) kernel for sports video matching and retrieval. Similar to SPM, the STPM is built to find the approximate correspondence of local feature points for a valid similarity metric between two video clips, by accumulating the matched feature points over a sequence of increasingly denser 3D grids in the spatial-temporal domain. Besides, Ni *et al.* [33] proposed a depth-layered spatial-temporal feature pooling method to utilize the depth information of STIPs, which shows boosted recognition performance compared to the conventional STIP BoW method for recognizing human daily activities in RGB-D videos. Moreover, recent research progress [20,34] demonstrates this to be an effective means of improving recognition performance, to combine multiple types of feature descriptors for utilizing their complementarity, and shows state-of-the-art performance on realistic human action recognition.

As shown in [10,24], the SPM/STPM kernel adopts *ad hoc* weights for combining the similarities in different grid granularity levels, which are inversely proportional to the cell width for penalizing the matching of local feature points found in larger cells. In spite of the success of SPM/STPM in many vision tasks, the fixed and predefined weights are not adaptive to the characteristics of training data. In this paper, we propose a new matching kernel to measure the similarity of two video clips, called the *generalized pyramid matching kernel* (GPMK), to leverage heterogeneous visual cues in multiple feature descriptor types and spatial-temporal grid granularity levels. Compared to the STPM kernel, our GPMK mainly has the following two advantages: (1) It extends standard STPM kernel into a more general form (*i.e.*, the STPM kernel presented in [10] can be deemed a special case of our GPMK.) and provides a more flexible fusion of multi-channel information for discrimination, by means of assigning independent weights on different channels of feature descriptor types and spatial-temporal grid granularity levels; (2) We present a simple, yet effective, approach to compute adaptive channel weights of GPMK based on the *kernel-target alignment* [35] from training data, which incorporates the prior knowledge and data-driven information of multiple channels in a principled way. In addition, we show that the proposed GPMK is a valid Mercer kernel, which is a desirable property for facilitating its usage in training kernel-based classifiers [36,37]. Thus, we apply it to the task of human action recognition in realistic videos via a kernel-based support vector machine (SVM) [36] classification framework. The experimental results on three challenging realistic video datasets (*i.e.*, Hollywood2 [14], Youtube [15] and HMDB51 [16]) validate the superiority of our GPMK *w.r.t.* traditional STPM kernel for human action recognition, and show higher classification accuracy than previous action recognition approaches in the literature. In summary, the main contributions of our method are listed below:

- We extend the STPM kernel and present a GPMK as a valid similarity metric between two video clips, which has a more general and flexible form to fuse heterogeneous information in multiple feature types and spatial-temporal grid levels.
- Instead of the predefined and fixed weights used in STPM, we propose a new method to compute the adaptive channel weights of GPMK based on the kernel-target alignment from training data.
- On the basis of the GPMK, we design a human action recognition system for realistic videos via a kernel-based SVM classification framework. The experimental results on three public datasets validate the superiority of our GPMK *w.r.t.* the STPM kernel for realistic human action recognition and outperform the state-of-the-art results in the literature.

The rest of this paper is organized as follows: In Section 2, we elaborate on the kernel-based SVM classification framework for human action recognition in videos. Section 3 introduces the construction of multi-channel BoW representation from local spatial-temporal features. After that, we formally present the GPMK, as well as the method for calculating adaptive channel weights in Section 4. In Section 5, we show experimental results and detailed analysis on the proposed GPMK. Finally, we conclude this paper in Section 6.

2. The Kernel-Based SVM Classification Framework

In this paper, we adopt the kernel-based SVM classification framework [36] for human action recognition in videos. Assuming there is a set of N video clips (denoted by $\{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_i, \cdots, \mathbf{x}_N\}$) used for training the classifier, let $y_i \in \{-1, 1\}$ denote the binary action class label of video clip $\mathbf{x}_i$. The values of 1 and -1 represent the positive and negative classes respectively, indicating whether the target human action appears in this video clip or not. We use $\mathbf{y} = [y_1, y_2, \cdots, y_i, \cdots, y_N]^{\mathrm{T}}$ to represent the label vector of all training samples accordingly. Then, we build a $N \times N$ kernel matrix $\mathbf{K}$ from training samples, in which the element of the i-th row and the j-th column is given by $\mathbf{K}_{ij} = \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$. In this paper, $\mathcal{K}(\cdot, \cdot)$ denotes the proposed GPMK, which is a valid kernel function for measuring the similarity between two video clips.

In the training stage, we learn a kernel SVM classifier by solving the following convex quadratic programming problem [36]:

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \tfrac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\widetilde{\mathbf{K}}\boldsymbol{\alpha} - \mathbf{e}^{\mathrm{T}}\boldsymbol{\alpha} \\ \text{s.t.} \quad & \mathbf{y}^{\mathrm{T}}\boldsymbol{\alpha} = 0, \quad \text{and} \quad 0 \leq \alpha_i \leq C, \quad i = 1, 2, \cdots, N \end{aligned} \tag{1}$$

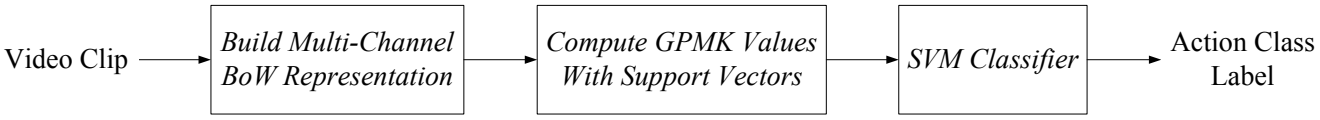
where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \cdots, \alpha_i, \cdots, \alpha_N]^{\mathrm{T}}$ denotes a coefficient vector whose entries are zero except for the support vector samples [36]. Meanwhile, $\mathbf{e}$ and C represent an N-dimensional unit vector and a predefined regularization parameter on training the SVM model, respectively. $\widetilde{\mathbf{K}}$ is a $N \times N$ label-augmented kernel matrix, where $\widetilde{\mathbf{K}}_{ij} = y_i y_j \mathbf{K}_{ij}$ for each pair of samples, i and j.

In the testing stage, we compute the classifier's score of a new video clip, $\mathbf{x}$, as follows:

$$f(\mathbf{x}) = \sum_{i \in sv} y_i \alpha_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i) + b \tag{2}$$

where sv and b denote the support vector set of training samples and a constant bias term, respectively. Thus, the class label of $\mathbf{x}$ can be identified via a sign function $y = \mathrm{sgn}(f(\mathbf{x}))$, where $y = 1$ for $f(\mathbf{x}) \geq 0$ and $y = -1$, otherwise. In Figure 1, we show the flowchart of kernel-based SVM classification for human action recognition in videos. For a test video clip, we first build a multi-channel BoW representation from the local spatial-temporal features extracted on it and, then, compute its GPMK values *w.r.t.* the learned support vectors of training samples. After that, we can classify this video clip by evaluating its score according to Equation (2). In the following part of this paper, we will elaborate on the multi-channel BoW representation and the GPMK in Sections 3 and 4, respectively.

Figure 1. The flowchart of our kernel-based human action recognition system.



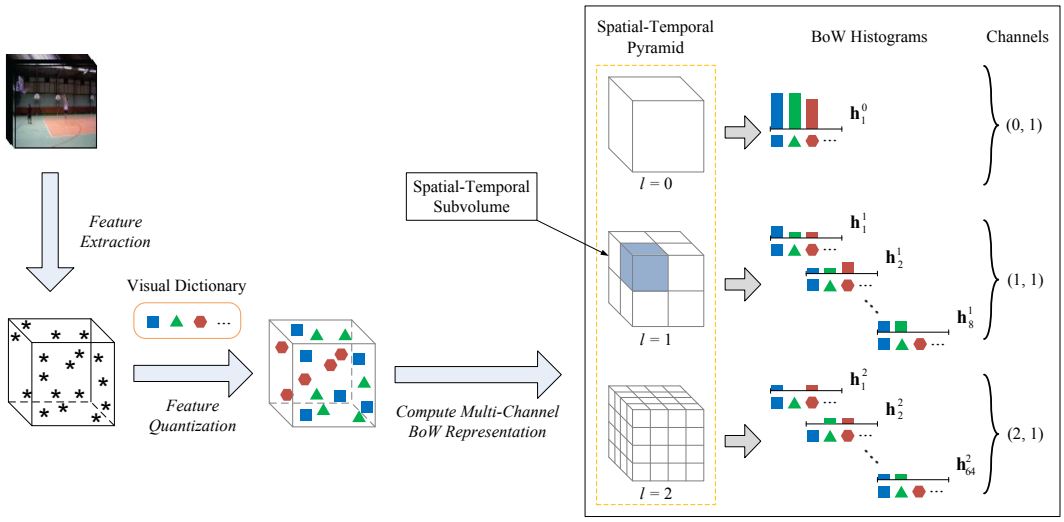
3. Multi-Channel BoW Representation for a Video Clip

Assuming there are Q feature descriptor types used in total, let $P_{\mathbf{x}}^{q}$ $(q \in \{1, 2, \cdots, Q\})$ denote a set of local spatial-temporal feature points extracted from video clip $\mathbf{x}$ for the q-th feature type. After that, given a dictionary with M visual words for each feature type, we discretize the descriptors of local feature points with corresponding visual word indices based on the vector quantization approach, as in [10,23,24]. Concretely, for every point, we compute the Euclidean distance between its feature descriptor vector and each of the visual words in the dictionary and map it to the visual word of minimum distance.

Similar to the STPM [10], let L denote the maximum grid granularity level of the spatial-temporal pyramid used. We partition a whole video clip into a sequence of evenly-divided 3D spatial-temporal grids with $(L + 1)$ different granularity levels. For each grid granularity level $l \in \{0, 1, \cdots, L\}$, $\mathbf{x}$ is uniformly partitioned by 2^l bins in each of the spatial-temporal dimensions, and thus, a total of $D_l = 2^{3l}$ subvolumes can be obtained. For each feature type $q \in \{1, 2, \cdots, Q\}$, we compute the occurrence frequency of visual words in the dictionary for different spatial-temporal subvolumes and obtain several BoW histograms [23] as the video representation used in GPMK. Let $\mathbf{h}_n^l(P_{\mathbf{x}}^q) = [h_{n,1}^l(P_{\mathbf{x}}^q), h_{n,2}^l(P_{\mathbf{x}}^q), \cdots, h_{n,m}^l(P_{\mathbf{x}}^q), \cdots, h_{n,M}^l(P_{\mathbf{x}}^q)]^{\mathrm{T}}$ denote an M-dimensional vector of the BoW histogram on quantized local feature points in the n-th subvolume of grid granularity level l, where $h_{n,m}^l(P_{\mathbf{x}}^q)$ is equal to the number of feature points belonging to the m-th visual word.

In this paper, we name the combination of grid granularity level l and feature type index q by a term *channel* (l, q) (We index the channel with a binary group notation for simplicity and intuition). Generally, it includes two aspects as follows: on the one hand, q indicates the type of feature descriptor used for capturing certain visual cues of local feature points extracted from the video clip; on the other hand, l corresponds to the grid granularity of encoding spatial-temporal distribution information for the quantized features. Thus, a video clip can be represented by a total of $Q \times (L + 1)$ BoW histograms in different feature descriptor types and spatial-temporal grid granularity levels, and we build a multi-channel BoW representation from local spatial-temporal features, which is used in the calculation of GPMK. Figure 2 illustrates the process of building the multi-channel BoW representation for a video clip.

Figure 2. Illustration of building the multi-channel bag of words (BoW) representation for a video clip. In this figure, the local feature points are represented by black stars, and the quantized features are shown with geometrical shape symbols in different colors (e.g., blue square, green triangle and red hexagon), each type of which indicates one visual word. A three-level spatial-temporal pyramid (*i.e.*, $L = 2$) is adopted for demonstration. For clarity, we show the case of using only one feature descriptor type (*i.e.*, $Q = 1$) as an example. For the case of multiple feature descriptor types used, the BoW histograms for the channels of each individual feature type can be computed in a similar way. (Note: the image is best viewed in color with magnification).



4. A Generalized Pyramid Matching Kernel with Adaptive Channel Weights

In this section, we first present the GPMK as a valid similarity metric between two video clips and, then, elaborate on the method of computing adaptive channel weights based on the kernel-target alignment [35].

4.1. The Generalized Pyramid Matching Kernel

In this paper, the GPMK is defined as a valid video similarity metric, for leveraging heterogeneous visual cues in multiple feature descriptor types and spatial-temporal grid granularity levels. Similar to STPM, we build the GPMK to find approximate correspondences of local feature points between two video clips, by accumulating the matched feature points over a sequence of increasingly denser 3D grids in the spatial-temporal domain.

Let $\mathbf{x}$ and $\mathbf{z}$ denote two video clips. At first, we compute the matching degree of these two video clips for each channel, respectively, which is defined by the total number of matched feature points over different visual words and spatial-temporal subvolumes in that channel. In practice, this can be calculated via a histogram intersection kernel (HIK) [38] function in Equation (3).

$$\mathcal{I}_q^l(\mathbf{x}, \mathbf{z}) = \sum_{n=1}^{D_l} \sum_{m=1}^{M} \min[h_{n,m}^l(P_{\mathbf{x}}^q), h_{n,m}^l(P_{\mathbf{z}}^q)] \tag{3}$$

The matching degree for channel (l, q), which is denoted by $\mathcal{I}_q^l$ in Equation (3), between $\mathbf{x}$ and $\mathbf{z}$ is equal to the sum of the minimum value on each bin of their BoW histograms $\mathbf{h}_n^l(P_\mathbf{x}^q)$ and $\mathbf{h}_n^l(P_\mathbf{z}^q)$ over all of the D_l subvolumes in level l.

After that, we construct the GPMK by combining the matching degree values over different channels, to build a unified similarity metric of video clips. Formally, it is defined by Equation (4), which is a weighted sum of $\mathcal{I}_q^l$ over all the channels.

$$\begin{aligned} &\mathcal{K}(\mathbf{x}, \mathbf{z}) = \textstyle\sum_{l=0}^{L} \sum_{q=1}^{Q} \omega_{(l,q)} \mathcal{I}_q^l(\mathbf{x}, \mathbf{z}) \\ \text{s.t.} \quad &\forall\ l \text{ and } q, \quad \omega_{(l,q)} \geq 0 \end{aligned} \tag{4}$$

where $\omega_{(l,q)}$ denotes the weight of channel (l, q) in GPMK.

As an extended form of the SPM/STPM kernel, the proposed GPMK is able to assign independent weight value for each individual channel (*i.e.*, the combination of feature descriptor type and grid granularity level) and provides a more flexible fusion of heterogeneous information for discrimination. As proven in [39,40], the HIK function of calculating $\mathcal{I}_q^l$ is positive definite. Besides, we constrain the weight $\omega_{(l,q)}$ to be non-negative for each channel. Thus, the resultant GPMK defined in Equation (4) is a conic sum (*i.e.*, linear combination with non-negative weights) [41] over a series of positive definite kernel functions on individual channels, such that it is a valid Mercer kernel, which facilitates the usage in training SVM classifiers [36,37]. Actually, according to [10], the STPM kernel is a special case of our GPMK that uses *ad hoc* weights by $\omega_{(l,q)} = \frac{1}{Q \cdot 2^L}$ for $l = 0$ and $\omega_{(l,q)} = \frac{1}{Q \cdot 2^{L-l+1}}$ for other levels $l \in 1, 2, \cdots, L$.

4.2. Adaptive Channel Weights Based on Kernel-Target Alignment

In this paper, rather than using fixed and predefined weights as in SPM/STPM, we compute adaptive channel weights for GPMK based on the *kernel-target alignment* [35]. Let $\mathbf{I}_q^l$ denote a $N \times N$ kernel matrix of channel (l, q) on training samples, in which the element of the i-th row and the j-th column is equal to $\mathcal{I}_q^l(\mathbf{x}_i, \mathbf{x}_j)$. Besides, we use the ground-truth class label vector, $\mathbf{y}$, to compute a target indicator matrix $\mathbf{Y} = \mathbf{y}\mathbf{y}^{\mathrm{T}}$, which presents an ideal discriminative characteristic on the classification of training samples. For each channel (l, q), we calculate the kernel-target alignment value, $s_{(l,q)}$, between $\mathbf{I}_q^l$ and $\mathbf{Y}$ on training data, which corresponds to the cosine of the angle between those two matrices [35]:

$$s_{(l,q)} = \frac{\langle \mathbf{I}_q^l, \mathbf{Y} \rangle_F}{\sqrt{\langle \mathbf{I}_q^l, \mathbf{I}_q^l \rangle_F \cdot \langle \mathbf{Y}, \mathbf{Y} \rangle_F}} = \frac{\langle \mathbf{I}_q^l, \mathbf{Y} \rangle_F}{N \sqrt{\langle \mathbf{I}_q^l, \mathbf{I}_q^l \rangle_F}} \tag{5}$$

In Equation (5), $\langle \mathbf{A}, \mathbf{B} \rangle_F = tr(\mathbf{A}^T \mathbf{B})$ represents the Frobenius inner product of two matrices, $\mathbf{A}$ and $\mathbf{B}$.

Actually, $s_{(l,q)}$ works as a simple and intuitive assessment on the discriminative power of channel (l, q) according to the training data. For instance, a larger alignment value for one channel implies that its kernel matrix is more similar to the target matrix, indicating higher discriminability for classification, and thus, a larger weight should be assigned to that channel. Based on the alignment values obtained

from training data for all individual channels, we compute an adaptive weight for each channel (l, q) in GPMK, via the following soft max activation function:

$$\omega_{(l,q)} = \frac{\pi_{(l,q)}\, e^{\beta s_{(l,q)}}}{\sum_{l'=0}^{L}\sum_{q'=1}^{Q} \pi_{(l',q')}\, e^{\beta s_{(l',q')}}} \quad \forall\ l = 0, 1, \cdots, L \text{ and } q = 1, 2, \cdots, Q \tag{6}$$

where $\pi_{(l,q)}$ is a non-negative prior term for channel (l, q) and β is a predefined parameter on the smoothness of soft max activation. In this paper, we set the value of $\pi_{(l,q)}$ to be consistent with the STPM kernel (*i.e.*, $\pi_{(l,q)} = \frac{1}{Q \cdot 2^L}$ for $l = 0$ and $\pi_{(l,q)} = \frac{1}{Q \cdot 2^{L-l+1}}$ for other levels $l = 1, 2, \cdots, L$). Thus, the proposed GPMK using weight values defined in Equation (6) would reduce to the standard STPM kernel when $\beta = 0$.

As observed in Equation (6), the channel weight of GPMK is determined based on two terms: on the one hand, the prior term $\pi_{(l,q)}$ represents the preference on the significance of channel (l, q); on the other hand, $e^{\beta s_{(l,q)}}$ is a data-driven term corresponding to the alignment value between the kernel matrix of channel (l, q) and the target matrix. Instead of the fixed and predefined weights used in SPM/STPM, we compute the channel weights of GPMK in a data-driven manner, which are adaptive to the characteristics of the training data. The parameter, β, controls the trade off between the prior term and the data-driven one. Moreover, a larger value of β highlights the channels with relatively higher alignment values via the soft max activation function. When the value of β increases, the GPMK will assign larger weights to more discriminative channels and shrink weight values of the channels with smaller alignment values through the L1-normalization of ω. Thus, our GPMK is able to adaptively select the channels with relatively higher discriminative power according to training data. Besides, observing that the weights defined in Equation (6) are always non-negative and subject to the constraints of Equation (4), the derived GPMK is indeed a valid Mercer kernel, as stated in Section 4.1.

5. Experiments

In this section, we evaluate the proposed GPMK on three public action datasets (*i.e.*, Hollywood2 [14], Youtube [15] and HMDB51 [16]) and show its superiority *w.r.t.* the STPM kernel [10] for human action recognition in realistic videos. Besides, we compare it with previous methods in the action classification literature and provide detailed analysis for the GPMK.

5.1. The Datasets

In this subsection, we briefly introduce the Hollywood2, Youtube and HMDB51 datasets used in our experiments. They provide realistic video clips for the action recognition benchmark under challenging conditions, such as camera motion, object appearance and pose, object scale, cluttered background, viewpoint and illumination variations, *etc*.

5.1.1. The Hollywood2 Dataset

This dataset is composed of video clips collected from 69 movies, which intends to provide a comprehensive benchmark for human action recognition in realistic scenarios. It consists of 12 human

action classes as follows: answering the phone, driving car, eating, fighting, getting out of car, hand shaking, hugging, kissing, running, sitting down, sitting up and standing up. Following common settings in the literature [14,18,20,42], we use the manually verified clean data in our experiments. There are a total of 1,707 video clips divided into a training subset (823 video clips) and a testing subset (884 video clips), which come from different movies. As in [14], we calculate the average precision (AP) of binary classification for each action class separately, and the overall performance is measured by the mean AP (mAP) over all classes.

5.1.2. The Youtube Dataset

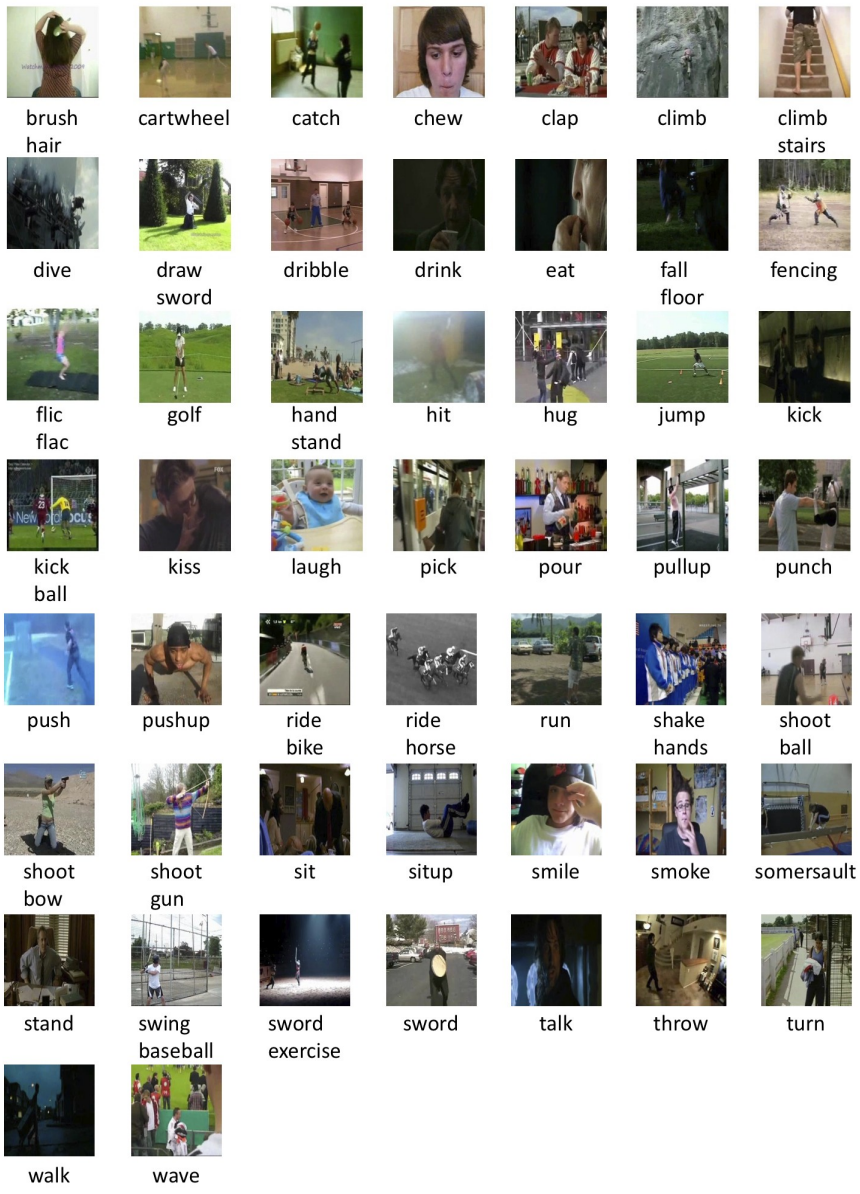
This dataset is collected from the Internet, which includes a total of 1,168 realistic video clips. It contains 11 human action classes: basketball shooting, biking/cycling, diving, golf swinging, horseback riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking and walking with a dog. Following the original protocol of [15], we adopt the predefined 25-fold leave-one-out cross validation for training and testing, and the performance is measured by the average of per-class classification accuracy over different folds. Figure 3 illustrates the sample frames from video clips for different human action classes in the Youtube dataset.

Figure 3. Illustration of the sample frames from video clips for all the 11 human action classes in the Youtube dataset [15]. (Note: the image is best viewed in color).

5.1.3. The HMDB51 Dataset

This dataset consists of 6,766 video clips distributed into 51 human action classes, which are collected from a variety of realistic video sources, such as movies and Internet videos. It contains the largest number of classes and video clips in the action classification literature. In experiments, we adopt the same three training/testing sample splits released by the authors of [16]. For each split, there are 70 training video clips and 30 testing ones per action class. Besides, we use the original non-stabilized version of video clips in our experiments. Following [16], the recognition performance is measured by the mean value of per-class classification accuracy over all three training/testing splits. Figure 4 illustrates the sample frames from video clips for different human action classes in the HMDB51 dataset.

Figure 4. Illustration of the sample frames from video clips for all 51 human action classes in the HMDB51 dataset [16]. (Note: the image is best viewed in color).



5.2. Experimental Settings

In the feature extraction step, we adopt a dense trajectory detector [20] with four different feature descriptor types (*i.e.*, motion boundary histogram [31], histogram of oriented gradients [17], histogram of optical flow [17] and trajectory shape descriptor [20]) to capture motion, appearance and shape cues of local spatial-temporal feature points. Furthermore, we use mean RGB values as an additional feature type to take advantage of the color information of densely extracted local feature points, and thus, there are a total of five kinds of feature descriptors (*i.e.*, $Q = 5$) used in our experiments. For the feature quantization step, we randomly select 100,000 samples of local feature points from training video clips and then, employ the K-means clustering algorithm [43] on their feature descriptors to build the visual dictionary. Specifically, the obtained cluster centers are used as the visual words in the dictionary. Following the common setting in the action classification literature [17,20,44], we set the number of visual words by $M = 4,000$ for each feature descriptor type.

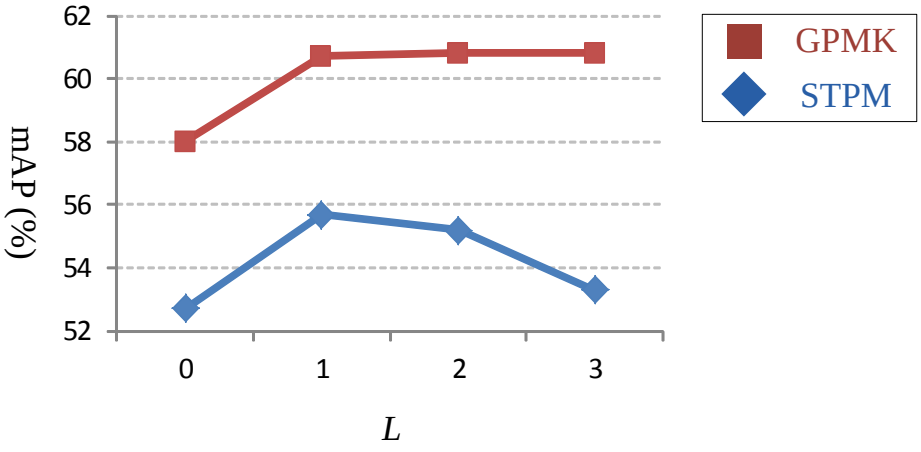
In building the GPMK, the value of parameter β in Equation (6) is set by 30 for Hollywood2, 20 for Youtube and 10 for HMDB51, respectively. In the classification step, we implement the SVM classifier by the LIBSVM code package [45] and adopt the "one-*vs*.-rest" criterion on multi-class discrimination for Youtube and HMDB51. Specifically, we train a binary SVM classifier for each action class individually, and then, the class label of the testing video is predicted as the one with the highest score output by the corresponding classifier. The regularization parameter, C, for training SVM classifiers is constantly set to 8 throughout our experiments.

5.3. Performance Evaluation

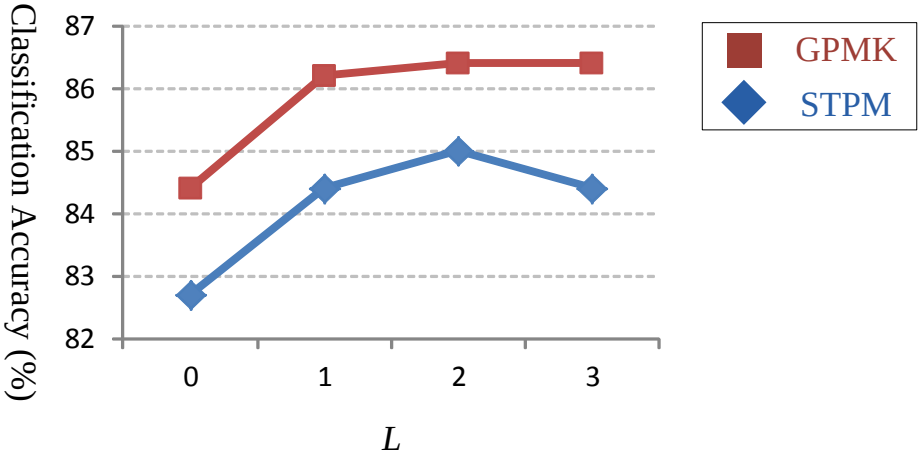
In this subsection, we evaluate recognition performance on the proposed GPMK and compare it with the STPM kernel. In practice, STPM is used as a baseline method in this paper, which actually corresponds to the special case of $\beta = 0$ for GPMK. Besides, we compare the result of our GPMK with previous methods in action classification literature.

Figure 5a–c show classification accuracy of the proposed GPMK on Hollywood2, Youtube and HMDB51 datasets, respectively, and compare it with the baseline method (*i.e.*, STPM). We can see that our GPMK consistently outperforms the STPM kernel for each number of different maximum grid granularity levels (*i.e.*, $L = 0, 1, 2, 3$) on all of the three datasets. Concretely, the performance gain of the GPMK *w.r.t.* STPM is 1.4%–2.0% for the Youtube dataset and 5.0%–7.5% for the Hollywood2 dataset. For the more challenging HMDB51 dataset, our GPMK can obtain 3.6%–4.6% improvement of classification accuracy *w.r.t.* the STPM kernel. The superiority of GPMK demonstrates that it is able to compute better weights by utilizing the data-driven information from training samples and, thus, leads to a more discriminative video similarity metric for kernel-based SVM classification. Besides, as shown in Tables 1–3, our method achieves superior classification accuracy compared to previous methods in action classification literature and reports the state-of-the-art results on these datasets.

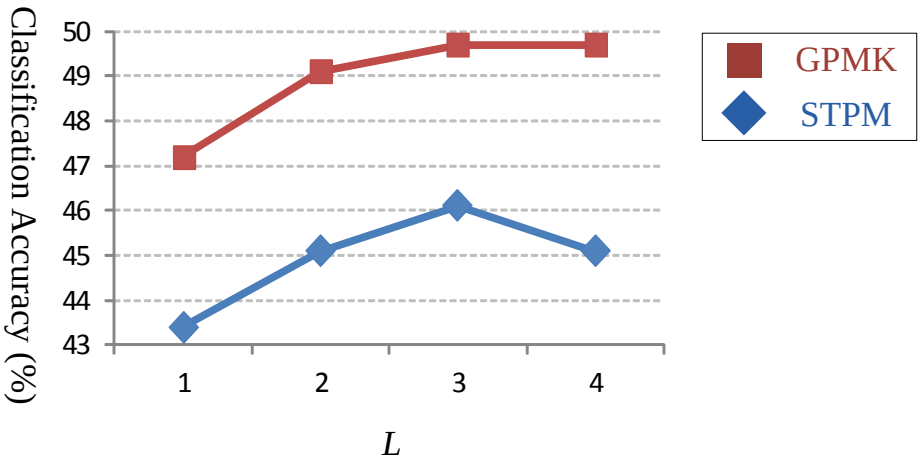
Figure 5. The recognition performance (shown on the Y-axis) *w.r.t.* different values of L (shown on the X-axis): (**a**) Hollywood2; (**b**) Youtube; (**c**) HMDB51. In each panel, the red square markers correspond to our generalized pyramid matching kernel (GPMK), and the blue diamond ones represent the spatial-temporal pyramid matching (STPM) kernel for comparison.



(a)



(b)



(c)

Table 1. Performance comparison of the proposed GPMK with previous methods in action classification literature (Hollywood2 [14]).

	mAP (%)
Our Method	**60.8**
Wang *et al.* [20]	59.9
Jiang *et al.* [34]	59.5
MIL-BoF [44]	48.73
L-MKL [46]	43.14
Le *et al.* [47]	53.3
Gilbert *et al.* [42]	50.9
Han *et al.* [18]	42.12
Marszalek *et al.* [14]	35.5

Table 2. Performance comparison of the proposed GPMK with previous methods in action classification literature (Youtube [15]).

	Classification Accuracy (%)
Our Method	**86.4**
Wang *et al.* [20]	85.4
MIL-BoF [44]	80.39
L-MKL [46]	77.91
Bhattacharya *et al.* [48]	76.5
Le *et al.* [47]	75.8
Human Postures [49]	77.8
Ikizler-Cinbis and Sclaroff [50]	75.21
Liu *et al.* [15]	71.2

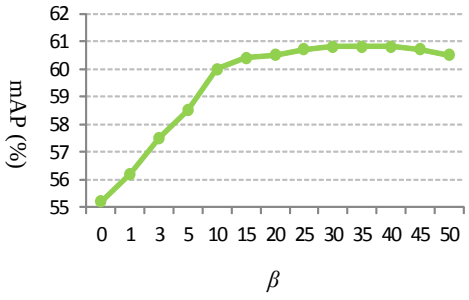
Table 3. Performance comparison of the proposed GPMK with previous methods in action classification literature (HMDB51 [16]).

	Classification Accuracy (%)
Our Method	**49.7**
Wang *et al.* [20]	48.3
Jiang *et al.* [34]	40.7
MIP [51]	29.17
MIL-BoF [44]	31.53
Action Bank [19]	26.9
Kuehne *et al.* [16]	22.83

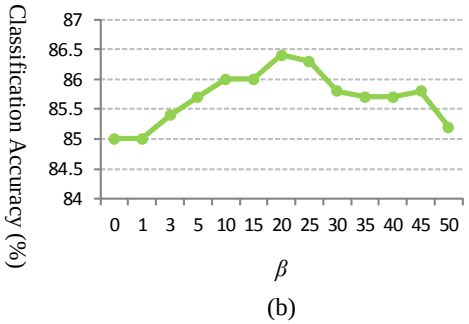
5.4. Analysis and Discussion

In this subsection, we provide detailed analysis and discussion on the proposed GPMK. As shown in Figure 5, we observe that the recognition performance achieves optimum with a 3-level pyramid (*i.e.*, $L = 2$) used for the GPMK, and thus, we set $L = 2$ for the discussion below. Figure 6 gives an empirical analysis on the effect of the smoothness parameter of soft max activation (*i.e.*, β) for our GPMK. We can see that there exists an intermediate value of β for achieving optimum performance. When the value of β is too small, the weights of GPMK tend to be almost fully determined by the prior term and cannot make good use of the data-driven cues. Particularly, the GPMK will reduce to the baseline method if $\beta = 0$. On the contrary, the recognition performance of GPMK deteriorates, as well, when using too large a value for β. This is mainly because an excessively large value of β causes the weights to sharply concentrate on the channels with the largest kernel-target alignment values, and thus, the GPMK tends to utilize only a very small part of information from all the channels. This observation supports our motivation for combining prior knowledge and data-driven information, as in Equation (6).

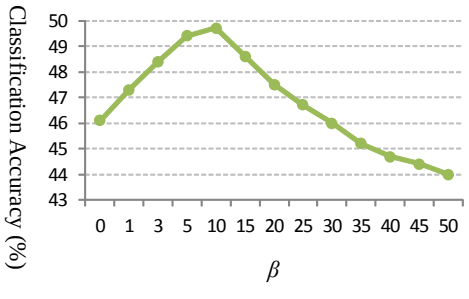
Figure 6. The recognition performance (shown on the Y-axis) *w.r.t.* different values of parameter β (shown on the X-axis) for GPMK: (**a**) Hollywood2; (**b**) Youtube; (**c**) HMDB51.



(a)



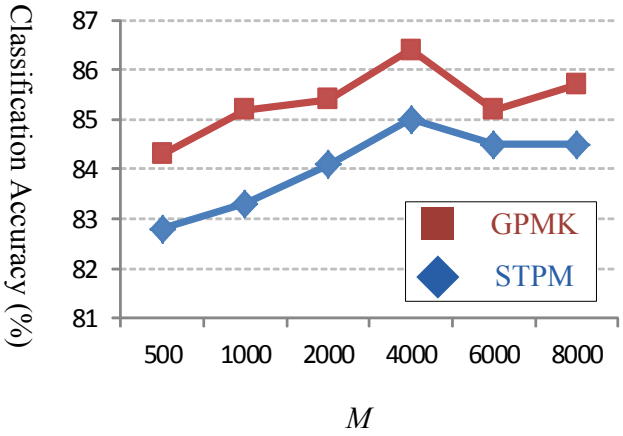
(b)



(c)

Moreover, we compare the GPMK with STPM for different numbers of visual words (*i.e.*, dictionary size M) used in experimental evaluation. As illustrated in Figure 7, our GPMK consistently outperforms the STPM kernel for each different value of M. This observation further validates the effectiveness of our method on computing adaptive channel weights in GPMK, and demonstrates its advantage *w.r.t.* the fixed and predefined weights used in STPM.

Figure 7. The classification accuracy (shown on the Y-axis) *w.r.t.* different values of M (shown on the X-axis) on Youtube dataset. The red square markers and the blue diamond ones correspond to our GPMK and the STPM kernel, respectively.



6. Conclusions

In this paper, we present a generalized pyramid matching kernel for human action recognition in realistic videos. It leverages heterogeneous visual cues in multiple feature descriptor types and spatial-temporal grid granularity levels, to build a valid video similarity metric for kernel-based SVM classification. Instead of the predefined and fixed weights used in SPM/STPM, we compute adaptive channel weights of GPMK according to the characteristic of training data. The experimental results on three public datasets validate the advantage of our GPMK *w.r.t.* the STPM kernel for human action recognition in realistic videos and show superior recognition performance compared to previous methods in literature.

Acknowledgments

This work was supported by NSFC (61071155,61271240), 973 National Program (2010CB731401, 2010CB731406) and STCSM (12DZ2272600).

Conflict of Interest

The authors declare no conflict of interest.

References

1. Turaga, P.K.; Chellappa, R.; Subrahmanian, V.S.; Udrea, O. Machine recognition of human activities: A survey. *IEEE Trans. Circuit. Syst. Video Technol.* **2008**, *18*, 1473–1488.
2. Xu, X.; Tang, J.; Zhang, X.; Liu, X.; Zhang, H.; Qiu, Y. Exploring techniques for vision based human activity recognition: Methods, systems, and evaluation. *Sensors* **2013**, *13*, 1635–1650.
3. Ke, S.R.; Thuc, H.L.U.; Lee, Y.J.; Hwang, J.N.; Yoo, J.H.; Choi, K.H. A review on video-based human activity recognition. *Computers* **2013**, *2*, 88–131.
4. Zhu, G.; Yang, M.; Yu, K.; Xu, W.; Gong, Y. Detecting Video Events Based on Action Recognition in Complex Scenes Using Spatio-Temporal Descriptor. In Proceedings of the ACM International Conference on Multimedia, Beijing, China, 19–23 October 2009; pp. 165–174.
5. Blunsden, S.; Fisher, R. The BEHAVE video dataset: Ground truthed video for multi-person behavior classification. *Ann. BMVA* **2010**, *4*, 1–12.
6. Park, J.Y.; Yi, J.H. Gesture recognition based interactive boxing game gesture recognition based interactive boxing game. *Int. J. Inf. Tech.* **2006**, *12*, 36–43.
7. Choi, J.; Cho, Y.; Han, T.; Yang, H.S. A View-Based Real-time Human Action Recognition System as an Interface for Human Computer Interaction. In Proceedings of International Conference on Virtual Systems and Multimedia, Brisbane, Australia, 23–26 September 2007; pp. 112–120.
8. Robertson, N.; Reid, I. A general method for human activity recognition in video. *Comput. Vis. Image Underst.* **2006**, *104*, 232–248.
9. Rodriguez, M.; Ahmed, J.; Shah, M. Action MACH: A Spatio-Temporal Maximum Average Correlation Height Filter for Action Recognition. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
10. Choi, J.; Jeon, W.J.; Lee, S.C. Spatio-Temporal Pyramid Matching for Sports Videos. In Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval, Vancouver, Canada, 30–31 October 2008; pp. 291–297.
11. Jones, S.; Shao, L.; Zhang, J.; Liu, Y. Relevance feedback for real-world human action retrieval. *Pattern Recogn. Lett.* **2012**, *33*, 446–452.
12. Schuldt, C.; Laptev, I.; Caputo, B. Recognizing Human Actions: A Local SVM Approach. In Proceedings of the 17th International Conference on Pattern Recognition, Cambridge, UK, 23–26 August 2004; pp. 32–36.
13. Gorelick, L.; Blank, M.; Shechtman, E.; Irani, M.; Basri, R. Actions as space-time shapes. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 2247–2253.
14. Marszałek, M.; Laptev, I.; Schmid, C. Actions in Context. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 2929–2936.
15. Liu, J.; Luo, J.; Shah, M. Recognizing Realistic Actions from Videos in the Wild. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1996–2003.

16. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A Large Video Database for Human Motion Recognition. In Proceedings of the IEEE International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2556–2563.
17. Laptev, I.; Marszalek, M.; Schmid, C.; Rozenfeld, B. Learning Realistic Human Actions from Movies. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
18. Han, D.; Bo, L.; Sminchisescu, C. Selection and Context for Action Recognition. In Proceedings of IEEE International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October 2009; pp. 1933–1940.
19. Sadanand, S.; Corso, J. Action Bank: A High-Level Representation of Activity in Video. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1234–1241.
20. Wang, H.; Klser, A.; Schmid, C.; Liu, C.L. Dense trajectories and motion boundary descriptors for action recognition. *Int. J. Comput. Vision* **2013**, *103*, 60–79.
21. Ramanan, D.; Forsyth, D.A. Automatic Annotation of Everyday Movements. In Proceedings of Advances in Neural Information Processing Systems, Vancouver and Whistler, BC, Canada, 8–13 December 2003.
22. Wang, L.; Cheng, L.; Thi, T.H.; Zhang, J. Human Action Recognition from Boosted Pose Estimation. In Proceedings of the International Conference on Digital Image Computing: Techniques and Applications, Sydney, Australia, 1–3 December 2010; pp. 308–313.
23. Fei-Fei, L.; Perona, P. A Bayesian Hierarchical Model for Learning Natural Scene Categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 524–531.
24. Lazebnik, S.; Schmid, C.; Ponce, J. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, New York, NY, USA, 17–22 June 2006; Volume 2, pp. 2169–2178.
25. Gemert, J.V.; Veenman, C.; Smeulders, A.; Geusebroek, J. Visual word ambiguity. *IEEE Trans. Pattern Anal. Mach. Intel.* **2010**, *32*, 1271–1283.
26. Yang, J.; Yu, K.; Gong, Y.; Huang, T. Linear Spatial Pyramid Matching Using Sparse Coding for Image Classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 1794–1801.
27. Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; Gong, Y. Locality-Constrained Linear Coding for Image Classification. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3360–3367.
28. Wang, H.; Ullah, M.M.; Klaser, A.; Laptev, I.; Schmid, C. Evaluation of Local Spatio-Temporal Features for Action Recognition. In Proceedings of British Machine Vision Conference, London, UK, 7–10 September 2009; pp. 124.1–124.11.
29. Laptev, I. On space-time interest points. *Int. J. Comput. Vision* **2005**, *64*, 107–123.

30. Dollár, P.; Rabaud, V.; Cottrell, G.; Belongie, S. Behavior Recognition via Sparse Spatio-Temporal Features. In Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, Beijing, China, 15–16 October 2005; pp. 65–72.
31. Dalal, N.; Triggs, B.; Schmid, C. Human Detection Using Oriented Histograms of Flow and Appearance. In Proceedings of the European Conference on Computer Vision, Graz, Austria, 7–13 May 2006; pp. 428–441.
32. Klaeser, A.; Marszalek, M.; Schmid, C. A Spatio-Temporal Descriptor Based on 3D-Gradients. In Proceedings of the British Machine Vision Conference, Leeds, UK, 1–4 September 2008; pp. 99.1–99.10.
33. Ni, B.; Wang, G.; Moulin, P. RGBD-HuDaAct: A Color-Depth Video Database for Human Daily Activity Recognition. In Proceedings of IEEE International Conference on Computer Vision Workshops, Barcelona, Spain, 6–13 November 2011; pp. 1147–1153.
34. Jiang, Y.G.; Dai, Q.; Xue, X.; Liu, W.; Ngo, C.W. Trajectory-Based Modeling of Human Actions With Motion Reference Points. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 425–438.
35. Cristianini, N.; Shawe-Taylor, J.; Elisseeff, A.; Kandola, J.S. On Kernel-Target Alignment. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 3–8 December 2001; pp. 367–373.
36. Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*; Cambridge University Press: Cambridge, UK, 2000.
37. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004.
38. Barla, A.; Odone, F.; Verri, A. Histogram Intersection Kernel for Image Classification. In Proceedings of the International Conference on Image Processing, Barcelona, Catalonia, Spain, 14–18 September 2003; pp. 513–516.
39. Maji, S.; Berg, A.C. Max-Margin Additive Classifiers for Detection. In Proceedings of IEEE International Conference on Computer Vision, Kyoto, Japan, 27 September–4 October, 2009; pp. 40–47.
40. Wu, J. A Fast Dual Method for HIK SVM Learning. In Proceedings of the European Conference on Computer Vision, Hersonissos, Heraklion, Crete, Greece, 5–11 September 2010; pp. 552–565.
41. Gonen, M.; Alpayd, E. Multiple kernel learning algorithms. *J. Mach. Learn. Res.* **2011**, *12*, 2211–2268.
42. Gilbert, A.; Illingworth, J.; Bowden, R. Action recognition using mined hierarchical compound features. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 883–897.
43. Hartigan, J.A.; Wong, M.A. A K-means clustering algorithm. *JSTOR: Appl. Stat.* **1979**, *28*, 100–108.
44. Sapienza, M.; Cuzzolin, F.; Torr, P.H. Learning Discriminative Space-Time Actions from Weakly Labelled Videos. In Proceedings of the British Machine Vision Conference, Guildford, Surrey, UK, 3–7 September 2012; pp. 123.1–123.12.

45. Chang, C.C.; Lin, C.J. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2011**, *2*, 27:1–27:27.
46. Song, Y.; Zheng, Y.T.; Tang, S.; Zhou, X.; Zhang, Y.; Lin, S.; Chua, T.S. Localized Multiple Kernel Learning for Realistic Human Action Recognition in Videos. *IEEE Trans. Circuit. Syst. Video Techn.* **2011**, *21*, 1193–1202.
47. Le, Q.; Zou, W.; Yeung, S.; Ng, A. Learning Hierarchical Invariant Spatio-Temporal Features for Action Recognition with Independent Subspace Analysis. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 3361–3368.
48. Bhattacharya, S.; Sukthankar, R.; Jin, R.; Shah, M. A Probabilistic Representation for Efficient Large Scale Visual Recognition Tasks. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 20–25 June 2011; pp. 2593–2600.
49. Brendel, W.; Todorovic, S. Activities as Time Series of Human Postures. In Proceedings of the European Conference on Computer Vision, Hersonissos, Heraklion, Crete, Greece, 5–11 September 2010; pp. 721–734.
50. Ikizler-Cinbis, N.; Sclaroff, S. Object, Scene and Actions: Combining Multiple Features for Human Action Recognition. In Proceedings of the European Conference on Computer Vision, Hersonissos, Heraklion, Crete, Greece, 5–11 September 2010; pp. 494–507.
51. Kliper-Gross, O.; Gurovich, Y.; Hassner, T.; Wolf, L. Motion Interchange Patterns for Action Recognition in Unconstrained Videos. In Proceedings of the European conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 256–269.