Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2010, Article ID 202768, 11 pages
doi:10.1155/2010/202768

Research Article

Novel Kernel-Based Recognizers of Human Actions

Somayeh Danafar, Alessandro Giusti, and Jürgen Schmidhuber

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), Galleria 2, 6928 Manno-Lugano, University of Lugano, Switzerland

Correspondence should be addressed to Somayeh Danafar, s.danafar@yahoo.com

Received 16 January 2010; Revised 18 March 2010; Accepted 8 April 2010

Academic Editor: ChangIck Kim

We study unsupervised and supervised recognition of human actions in video sequences. The videos are represented by probability distributions and then meaningfully compared in a probabilistic framework. We introduce two novel approaches outperforming state-of-the-art algorithms when tested on the KTH and Weizmann public datasets: an *unsupervised* nonparametric kernel-based method exploiting the Maximum Mean Discrepancy test statistic; and a *supervised* method based on Support Vector Machine with a characteristic kernel specifically tailored to histogram-based information.

1. Introduction

Huge video archives require advanced video analysis to automatically interpret, understand, and summarize the semantics of video contents. In this paper we focus on localizing and categorizing different human actions in surveillance videos.

The task is challenging as visual perceptions of such events are very high-dimensional, and huge intraclass variations are common due to view point changes, camera motion, occlusions, clothing, cluttered background, geometric, and photometric object distortions.

A large amount of literature deals with this problem, by applying a variety of different techniques, which we briefly review in Section 2; most approaches however share a common high-level structure:

(i) extraction of features from the video data,

(ii) if necessary, dimensionality reduction of feature vectors, by means of techniques such as PCA,

(iii) classification of the sequence.

Our main contributions are new techniques for the third step, classification.

In our experimental evaluation we consider two different state-of-the-art feature descriptors, which have been described in action recognition systems providing top-tier results on the publicly available KTH [1, 2] and Weizmann [3] datasets. By using our proposed classification algorithm with such features, we manage to further improve classification results on the same datasets.

In order to be sufficiently powerful to descriptively represent video content, such features are high dimensional. This is commonly handled by using kernel-based methods, which allow one to perform classification implicitly in a reduced space.

In this framework, we deal with two problems:

(i) unsupervised clustering of sequences from unlabeled data, given the desired number of clusters;

(ii) supervised classification of new input sequences, given a set of labeled training sequences.

In the unsupervised case, the core idea is to represent each sequence as a probability distribution: if two probability distributions are similar enough, the corresponding video sequences are expected to represent the same action (see Figure 1).

In order to enable meaningful comparisons between probability distributions, such distributions are embedded in a high-dimensional Reproducing Kernel Hilbert Space (RKHS) by means of *characteristic kernels*, which enable injective embedding of probabilities [4–7]. The distance between mapped distributions is known as Maximum Mean Discrepancy (MMD) [8, 9], whose well-defined application is homogeneity testing.

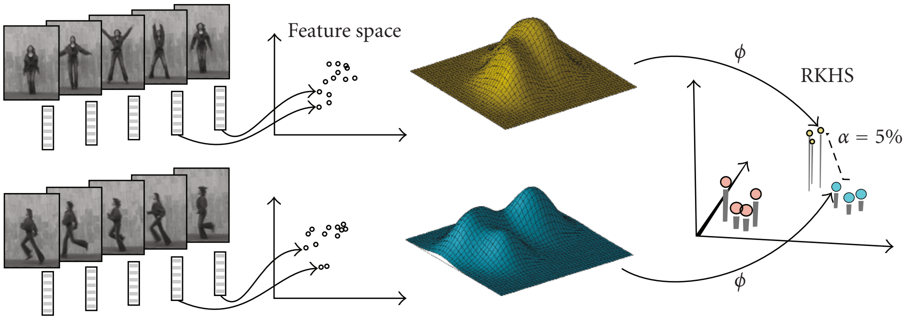


Figure 1: Unsupervised action recognition is performed by computing a feature vector for each frame of the sequence (left), as described in Section 6. A nonparametric probability distribution on the feature space is computed for each sequence (center); different sequences are compared by mapping the corresponding probability distributions to an RKHS by means of ϕ (right): in particular, sequences representing the same actions are clustered by using MMD as a distance metric.

(i) The first main contribution of this paper is the novel use of MMD as a homogeneity test for unsupervised action recognition (see Figure 1). Its encouraging performance, exceeding the best results in the literature, suggests that our classification technique is well suited to action recognition problems, and manages to capture differences between different classes while being robust to the significant appearance variations in the provided datasets. This is in accordance to several works in the literature, where MMD has been successfully used for unsupervised tasks in several different applications (see Section 2).

(ii) The second main contribution is in the supervised case: we use an SVM-based approach (see Figure 2), with a novel characteristic kernel specifically tailored for histogram-based data. Also in this context, we provide experimental evidence that selecting an appropriate kernel leads to significant performance gains.

By representing video sequences by means of probability distributions of feature vectors associated to video frames, we implicitly disregard frame ordering; such property is shared by several other approaches exploiting bag-of-features techniques [10, 11], and allows us to bypass the problem of determining the initial or final times of an action, while at the same time taking advantage of the action periodicity.

We review related literature in Section 2. In Section 3 we illustrate Maximum Mean Discrepancy, which is the core of our unsupervised method, and review the definition of characteristic kernels. Next we address the case of characteristic kernels which are defined for Abelian semigroups in Section 4. It gives us a characteristic kernel which is proper for histogram-based feature descriptors that we use in our supervised method. We discuss about the general framework of our unsupervised and supervised approaches in Section 5. In Section 6, we provide a brief overview to the feature extraction approaches that we use for our experimental validation, which is described in Section 7 using KTH and Weizmann datasets. In Section 7, we also discuss computational cost. Lastly, we draw conclusions and discuss future works in Section 8.

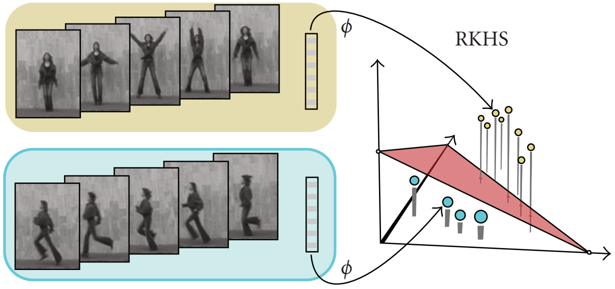


Figure 2: For supervised action recognition, for each sequence we compute a single feature vector (left) representing histogram-based data, as described in Section 6. Such feature vectors are then mapped to an RKHS by using a novel kernel which is both characteristic and appropriates for said representation (right). An SVD classifier is learnt from labeled training sequences in order to classify different actions.

2. Related Works

A large amount of different approaches have been proposed so far for action recognition (a recent review is given in [12]). We provide a broad classification in the following.

2.1. Features for Action Recognition. *Shape-based* approaches attempt to extract silhouettes of actors and then recognize the actions by analyzing such data [3, 13–16]. One inherent disadvantage of this class of techniques is that they can not capture the internal motion of the object within the silhouette region. More importantly, even state-of-the-art background subtraction techniques are unable to reliably recover precise silhouettes, especially in dynamic environments, which reduces the robustness of techniques in this class.

Flow-based techniques estimate the optical flow field between adjacent frames and use such features for action recognition, and provide the important advantage of requiring no background subtraction. A pioneering algorithm in this category was proposed by Efros [17]. They reported their results on a database of images taken at distance. Shechtman and Irani [18] use a template matching approach to correlate the flow consistency between the template and the video. Danafar and Gheissari [19] proposed an optical-flow-based algorithm which has the advantages of both holistic (look at human body as whole) and body-part-based approaches. This is one of the two descriptors used in this paper, and is outlined in Section 6. Jhuang et al. [20] extract dense local motion information with a set of flow filters. The responses are pooled locally, and converted to higher-level responses using complex learned templates. These templates are pooled again, and fed into a discriminative classifier.

In order to design features robust to changes in camera view and variability in the speed of actions, some researchers proposed space-time interest point features [1, 2, 10]. Dollár et al. [21] present a spatiotemporal interest point detector based on 1D-Gabor filters, which identifies regions with sudden or periodic intensity changes in time. Thereafter for each 3D interest region, optical flow descriptors are obtained. A fixed set of 3D visual words is compared with a histogram of a new sequence of visual words by a nearest neighbor approach. Ke et al. [22] also presented a new spatiotemporal shape and flow correlation algorithm for action recognition which works on oversegmented videos and does not require background subtraction.

Using both form and flow features simultaneously is also suggested in the seminal work of Giese and Poggio [23], which describes the strategy of biological systems: form and motion are processed simultaneously but independently in two separate pathways. However in their paper. The implementation of such system is designed for simple, schematic stimuli.

The approach is taken further by Schindler and Van Gool [24], which investigates the detection of actions from very short sequences called snippets. The motion pathway extracts optic flow at different scales, directions, and speeds. In the form pathway, they apply Gabor filter at multiple orientations and scales. In both pathways, the filter responses are MAX-pooled, and computed to a set of learned templates. The similarities from both pathways are concatenated to a feature vector and classified with a bank of linear classifiers by SVM. In our approach, we use such powerful feature descriptor, computed on each pair of frames independently, as the input of our classification algorithm.

2.2. Classification for Action Recognition. Many classification techniques are proposed in literature, both supervised and unsupervised.

In [25], the authors propose compound features that are assembled from simple 2D corners in both space and time. Compound features are learned in a weakly-supervised approach using a data mining algorithm. Several researchers have explored unsupervised methods for motion analysis. Hoey [26] applies a hierarchical dynamic Bayesian network model to recognize facial expressions in an unsupervised manner. Zhong et al. [27] have proposed an unsupervised approach to detect unusual activity in video sequences. A simple descriptor vector per each frame is considered and video is clustered by looking at co-occurrences of motion and appearance patterns. Their method identifies spatially isolated clusters as unusual activity. In [28], the authors detect abnormal activities by means of the multi-observation Hidden Markov Model and spectral clustering to unsupervised training of behavior models. Boiman and Irani [29] explain a video sequence using patches from a database; as dense sampling of the patches is necessary in their approach, the resulting algorithm is very time consuming and unpractical for action recognition. Wang et al. [30] propose to use an unsupervised learning approach to discover the set of action classes present in a large collection of training images. Thereafter, these action classes are used to label test images. The features are based on the coarse shape of human figures and the distance between a pair of images is computed using a linear programming relaxation technique. Spectral clustering is performed using the resulting distances. Niebles et al. [11] present an unsupervised learning method for human action categories. Their algorithm automatically learns the probability distribution of the spatiotemporal words that each corresponds to an action category, and builds a model for each class. This is achieved by using latent topic models such as probabilistic Latent Semantic analysis (pLSA) model and Latent Dirichlet Allocation (LDA).

Many researchers use supervised and discriminative approaches for the classification stage, particularly with Support Vector Machines with an appropriate kernel according to feature descriptors [2, 19, 24, 31, 32]. Other approaches represent videos by using sparse spatiotemporal words, then summarized in a histogram. In such approaches, the temporal order of frames is disregarded, which is also shared in our approach. Nowozin et al. [31] propose a sequential representation which retains the temporal order. They introduce a discriminative subsequent mining to find optimal discriminative subsequent patterns, and extend the prefix span subsequence mining algorithm [33] in combination with LPBoost [34].

Maximum Mean Discrepancy as a statistical test has application in variety of areas. For instance, in bioinformatics we might wish to find whether different procedures in different labs on the same tissue obtain different DNA microarry data [35]. In database attribute matching has been used for merging heterogeneous databases [8]. In speaker verification, such test can be used to identify the correspondence between a speech sample to a person for whom previously recorded speech is available [36]. In this paper we propose a novel use of MMD as an unsupervised action recognition method.

3. The Maximum Mean Discrepancy

In this section we briefly recall the theoretical foundations of MMD: in Section 5, we show how it is employed in our context.

Recent studies [5, 6, 8] have shown that mapping random variables into a suitable reproducing kernel Hilbert space (RKHS) gives a powerful and straightforward method of dealing with higher order statistics of the variables. The idea behind this is to do linear statistics in RKHS and derive its meaning in the original space. One basic statistic on Euclidean space is the *mean*. By embedding the distributions to RKHS, the corresponding factor is the *mean element*, which was introduced by Gretton et al. [8, 9]. The distance between mapped mean elements is known as Maximum Mean Discrepancy (MMD). One well-defined application of MMD is for homogeneity testing or for the two sample test. The two sample problem tests whether two probability measures P and Q coincide or not.

Definition 1. Let $\mathcal{F}$ be an RKHS on the separable metric space $\mathcal{X}$, with a continuous feature mapping $\phi(x) \in \mathcal{F}$ for each $x \in \mathcal{X}$. The inner product between feature mappings is given by the positive definite kernel function $k(x, x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$. We assume that the kernel k is bounded. Let $\mathcal{P}$ be the set of Borel probability measures on $\mathcal{X}$.

Following [4, 8, 9], we define the mapping to $\mathcal{F}$ of $P \in \mathcal{P}$ as the expectation of $\phi(x)$ with respect to P (i.e., the mean element μ_P):

$$\begin{aligned} \mu_P : \mathcal{P} &\longrightarrow \mathcal{F}, \\ P &\longmapsto \int_{\mathcal{X}} \Phi(x) dP. \end{aligned} \tag{1}$$

The definition of MMD is explained in the following theorems [8, 9].

Theorem 2. *Let P and Q be two Borel probability measures defined on $\mathcal{X}$. Then $P = Q$ if and only if* $\text{MMD}[P, Q] = 0$. *Let x, x' be independent random variables drawn according to P, and y, y' be independent and drawn according to Q, and let x be independent of y. Then,*

$$\begin{aligned} &\text{MMD}[P, Q] \\ &\quad := \left\|\mu_P - \mu_Q\right\|_{\mathcal{H}} = \left\|E_P[k(x, \cdot)] - E_Q[k(y, \cdot)]\right\|_{\mathcal{H}} \\ &\quad = \left(E_{x,x'}(k(x, x')) + E_{y,y'}k(y, y') - 2E_{x,y}k(x, y)\right)^{1/2}. \end{aligned} \tag{2}$$

In practice, because we do not have access to the population of distributions P and Q, we compare two sets of data which are drawn from the populations. The homogeneity test becomes a problem of testing whether two samples of random variables are generated from the same distribution. MMD_u, the unbiased empirical estimation of the MMD is defined as follows.

Definition 3. Given observations $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_n\}$, drawn independently and identically distributed from P and Q, respectively, the unbiased estimate of MMD is the one-sample U-statistic:

$$\text{MMD}_u^2 := \frac{1}{m(m-1)} \sum_{i \neq j}^{m} h\left(z_i, z_j\right), \tag{3}$$

where $z_i := (x_i, y_i)$, $h(z_i, z_j) := k(x_i, x_j) + k(y_i, y_j) - k(x_i, y_i) - k(x_j, y_i)$, and m is the sample size.

The *biased* estimate MMD_b^2 is achieved by replacing the U-statistic in the above equation with a V-statistic (then the sum includes the term $i = j$).

In the two sample test, we require both a measure of distance between probabilities and a notion of whether this distance is statistically significant. The former is given in Theorem 2. For the latter, we give an expression for the asymptotic distribution of this distance measure, from which a significance threshold may be obtained. More precisely, we conduct a hypothesis test with null hypothesis $\mathcal{H}_0$ defined as $P = Q$, and alternative hypothesis $\mathcal{H}_1$ as $P \neq Q$. We must therefore specify a threshold that the empirical MMD will exceed with small probability when $P = Q$.

Theorem 4. *Let P and Q be two Borel probability measures defined on $\mathcal{X}$. Let $X := \{x_1, \ldots, x_m\}$ and $Y := \{y_1, \ldots, y_n\}$ be observations which are drawn independently and identically from P and Q, respectively. Let us assume $0 \leq k(x, y) \leq K$. Then*

$$\begin{aligned} \Pr\Bigg\{ &\left|\text{MMD}_b[X, Y] - \text{MMD}[P, Q]\right| \\ &> 2\left(\left(\frac{K}{m}\right)^{1/2} + \left(\frac{K}{n}\right)^{1/2}\right) + \epsilon \Bigg\} \leq 2 \exp\left(\frac{-\epsilon^2 mn}{2K(m+n)}\right). \end{aligned} \tag{4}$$

In [8] the proof of Theorem 4 has been shown by means of so called *Rademacher average*. We accept the null hypothesis $P = Q$ if the value of $\text{MMD}_b(P, Q)$ satisfies the inequality in Corollary 5 and reject the null hypothesis if not.

Corollary 5. *A hypothesis test of level α for the null hypothesis $P = Q$ (that is for* $\text{MMD}[P, Q] = 0$*) has the acceptance region*

$$\text{MMD}_b[X, Y] < \sqrt{\frac{2K}{m}} \left(1 + \sqrt{2 \log \alpha^{-1}}\right), \tag{5}$$

where α is the user-defined significance threshold (confidence interval) for test statistic.

In practice we used the looser significance threshold that is defined in Corollary 5. Empirically to estimate the boundary, the bootstrap method of Gretton et al. [8, 25] on the aggregated data is used. For theoretical point of view we elaborate on a tighter significance threshold of our two-sample test which is obtained by an expression of the asymptotic distribution. The following theorem explains that the unbiased empirical version of MMD asymptotically converges to the population value of MMD and obtains the threshold.

Theorem 6. *We assume $E(h^2) < \infty$. Under $\mathcal{H}_1$,* MMD_u^2 *converges in distribution to a Gaussian according to*

$$m^{1/2}\left(\text{MMD}_u^2 - \text{MMD}^2[P, Q]\right) \xrightarrow{D} \mathcal{N}\left(0, \sigma_u^2\right), \tag{6}$$

where $\sigma_u^2 = 4(E_z[(E_{z'}h(z, z'))^2] - [E_{zz'}(h(z, z'))]^2)$, uniformly at rate $1/\sqrt{m}$. Under $\mathcal{H}_0$, the U-statistic is degenerate, meaning $E_{z'}h(z, z') = 0$. In this case, MMD_u^2 *converges in distribution according to*

$$m\text{MMD}_u^2 \xrightarrow{D} \sum_{l=1}^{\infty} \lambda_l \left[z_l^2 - 2\right], \tag{7}$$

where$z_l \sim \mathcal{N}(0,2)$ *i.i.d.,* λ_i*are the solutions to the eigenvalue equation*

$$\int_{\mathcal{X}} \hat{k}(x,x')\psi_i(x)dp(x) = \lambda_i\psi_i(x'), \tag{8}$$

and $\hat{k}(x_i,x_j) := k(x_i,x_j) - E_x k(x_i,x) - E_x k(x,x_j) + E_{x,x'}k(x,x')$*is the centered RKHSkernel.*

The goal is to determine whether the empirical test statistic MMD_u^2 is so large to be outside the $1-\alpha$ quantile of the null distribution z_l (consistency of the resulting test is guaranteed by the form of the distribution under $\mathcal{H}_1$). One way to estimate this quantile is using the bootstrap on the aggregated data [8, 9].

Clearly the quality of the MMD as a statistic depends on the richness of RKHS space $\mathcal{H}$ which is defined by a measurable kernel k. A set of kernels is called *characteristic kernels*, introduced in [4, 5] gives an RKHS for which probabilities have unique images. The necessary and sufficient condition for a kernel to be characteristic is expressed in following lemma.

Lemma 7. *Let* $(\mathcal{X},\mathcal{B})$ *be a measurable space, k be a measurable positive definite kernel on* $\mathcal{X}$*, and* $\mathcal{H}$ *be the associated RKHS. Also let* $\mathbb{R}$ *be an RKHS, then k is characteristic if and only if* $\mathcal{H}+\mathbb{R}$ *is dense in* $L^2(P)$ *for every probability P on* $(\mathcal{X},\mathcal{B})$*.*

The definition of a characteristic kernel generalizes the well-known property of the characteristic functions which uniquely determines a Borel probability measure. The Gaussian RBF kernel $k(x,y) = \exp(-\|x-y\|^2/\sigma^2)$ is a famous example of a characteristic kernel on the entire $\mathbb{R}^m$. We use this kernel in the present work for unsupervised action recognition, whereas in the supervised case we introduce a different characteristic kernel in the following section.

4. Characteristic Kernels on Abelian Semigroups

Our supervised action recognition approach, outlined in Section 5, is based on SVM. The crucial condition that a kernel should satisfy to be suitable for SVM is to be positive definite, meaning that the SVM problem is convex, and hence that the solution of its objective function is unique. Positive definite kernels are defined as following.

Definition 8. Let $\mathcal{X}$ be a nonempty set. A function $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ is called positive definite kernel if and only if it is symmetric (i.e., $k(x,x') = k(x',x)$) for all $x, x' \in \mathcal{X}$ and if

$$\sum_{i,j=1}^{n} c_i c_j k\left(x_i,x_j\right) \geq 0, \quad \forall c \in \mathbb{R}. \tag{9}$$

In kernel-based methods like SVM, the choice of kernel is extremely important for the classification performance. As we have histogram-based feature vectors, we discuss here on positive definite kernels which are proper for histograms and then on characteristic kernels tailored on histogram-based information. We consider *Histogram Intersection (HI)* kernel as a positive definite kernel. HI has been first introduced in computer vision by Swain and Ballard in [37]:

$$k_{\mathrm{HI}}(a,b) = \sum_{i=1}^{n}\min(a_i,b_i), \tag{10}$$

where $x = (a_1,\ldots,a_n)$ and $b = (b_1,\ldots,b_n)$ are two n bins histograms (in $\mathbb{R}^n$). This kernel was successfully used as a similarity measure for image retrieval and recognition tasks [38, 39]. In [38] they proved that for histograms of the same size with integer values, k_{HI} is a positive definite kernel.

In [39] the *Generalized Histogram Intersection kernel* was introduced as a positive-definite kernel:

$$k_{\mathrm{GHI}}(a,b) = \sum_{i=1}^{n}\min\left(\left|a_i^{\beta}\right|,\left|b_i^{\beta}\right|\right), \quad (a,b)\in\mathcal{X}\times\mathcal{X}, \tag{11}$$

where $\beta \geq 0$ and $\mathcal{X} \in \mathbb{R}$. If we set $\beta = 1$, the k_{HI} is a special case of k_{GHI} and is a positive definite kernel for absolute real values.

Characteristic kernels have positive definite property and have been shown to be more discriminative, because they can take higher order statistics into account. For instance, in [40] Fukumizu et al. showed by optimizing kernel mappings one can find the most predictive subspace in regression. We verify this in practice in Section 7, where we show that the characteristic kernel $k(a,b) = e^{-\beta\sum_{i=1} n\sqrt{a_i+b_i}}$ provides significantly better performance than an HI kernel. Previously, characteristic kernel has been defined on $\mathbb{R}^n$ spaces. However, the kernel should be chosen according to the nature of the available data. In our supervised recognition case, just like in many other computer vision tasks, features are histogram-based, and are not naturally represented in the $\mathbb{R}^n$ space.

Therefore, we are going to investigate whether characteristic kernels can be defined on spaces besides $\mathbb{R}^n$. Several such domains constitute topological groups or semigroups; this is relevant in our context, as histograms are examples of Abelian semigroups.

Fukumizu et al. [6] introduced characteristic kernels on groups and semigroups by establishing some conditions. In this section we first recall the Bochner theorem which characterizes a set of continuous shift-invariant positive-definite kernels on $\mathbb{R}^n$ by the Fourier transform. Thereafter we bring the related theorems, which define characteristic kernels for Abelian semigroups and it is achieved based on Laplace transform in the Bochner theorem. The purpose here is to introduce a class of characteristic kernels for histograms that are examples of Abelian semigroups.

Theorem 9 (Bochner). *Let* $\phi : \mathbb{R}^n \to \mathbb{C}$ *be a bounded continuous function.* ϕ *is positive definite if and only if there is a unique finite nonnegative Borel measure* Λ *on* $\mathbb{R}^n$ *such that*

$$\phi(x) = \int_{\mathbb{R}^n} e^{\sqrt{-1}x^T\omega}d\Lambda(\omega), \tag{12}$$

where $\omega \in \mathbb{R}^n$*.*

Before explanation of the related theorem on semigroups we briefly review the definition of semigroups.

Definition 10. A semigroup $(S, \circ)$ is a nonempty set S equipped with an operation $\circ$ that satisfies the associative law:

$$(x \circ y) \circ z = x \circ (y \circ z), \tag{13}$$

for any $x, y, z \in S$. A semigroup $(S, \circ)$ is said to be Abelian if the operation is commutative, that is,

$$x \circ y = y \circ x, \tag{14}$$

for any $x, y \in S$.

Theorems 11 and 12 [6] obtain necessary and sufficient conditions for tailored kernels on Abelian semigroups $(\mathbb{R}^n_+, +)$.

Theorem 11. *Let $\phi : \mathbb{R}^n_+ \to \mathbb{C}$ be a bounded continuous function on $\mathbb{R}^n_+$. ϕ is positive definite if and only if there exists a unique nonnegative measure $\Lambda \in M(\mathbb{R}^n_+)$ such that*

$$\phi(x) = \int_{\mathbb{R}^n_+} e^{-\sum_{i=1}^n t_i x_i} d\Lambda(t) \quad (\forall x \in \mathbb{R}^n_+). \tag{15}$$

Based on the above theorem, we have the following sufficient condition of characteristic property.

Theorem 12. *Let ϕ be a positive definite function given equation in Theorem 11. If* $\operatorname{supp}(\Lambda) = \mathbb{R}^n_+$*, then the positive definite kernel $k(x, y) = \phi(x + y)$ is characteristic.*

As histograms represent an example of Abelian Semigroups, we take advantages of Theorems 11 and 12, and define this following *Histogram Characteristic* kernel.

Histogram Characteristic Kernel. Let $a = (a_i)^n_{i=1}$ and $b = (b_i)^n_{i=1}, (a_i \geq 0, b_i \geq 0)$ be nonnegative measures on n points, and $t \in \mathbb{R}^n_+$. k_{HC} is defined as

$$\Lambda = t^{-3/2} e^{-\beta^2/(4t)} \quad (\beta > 0) : \qquad k_{\mathrm{HC}}(a, b) = e^{-\beta \sum_{i=1}^n \sqrt{a_i + b_i}}. \tag{16}$$

Our proposed HC kernel provides significantly better performance than both the HI kernel (which is just positive definite, and not characteristic), and the Gaussian kernel (which is characteristic but not tailored on histogram-based information).

5. Unsupervised and Supervised Action Recognition

In this paper, we are applying the theoretical findings reported in the previous sections to two different problems: unsupervised and supervised action recognition.

In the *unsupervised* case, we aim at clustering unlabeled sequences belonging to the same action, assuming that the number of clusters is known. In this problem we use MMD with a gaussian kernel, as introduced in Theorem 4. σ is automatically determined, in such a way to return the required number of clusters. We considered the significance level, α, of MMD as a two-sample test equal to 0.05. The reported results are percentage of acceptance rate in 1000 times running the MMD. Clusters are found by pairwise comparisons (two-sample test) of distributions corresponding to sequences: two sequences belong to the same cluster if and only if the MMD is close enough to 0 (the threshold is computed as in Theorem 4 and Corollary 5). For each cluster, a single representative distribution is then chosen. Thereafter, a new sequence can be classified by comparing with the same approach its related probability distribution, to the representative distribution of each of the clusters (see Figure 1). For *supervised* action recognition, we use as a learning algorithm an SVM with the characteristic kernel, introduced in Example. The dataset is divided in three parts: training, testing, and validation. The validation data is first used in order to tune the β parameter of the kernel with a leave-one-out cross-validation procedure. According to the results of cross validation procedure β tuned as 0.001 for HC kernel and 1 for GHI kernel (which obtains the HI kernel). Then we use the training data in order to obtain support vectors which define the discriminative classifier. Lastly, the testing data is processed in order to evaluate the performance of the classifier (prediction, see Figure 2).

6. Feature Extraction Approaches

We evaluated our classification approach with two state of the art feature descriptors, which we will refer to as F1 and F2 in the following. They have been described in recent literature and shown to have excellent performance on the action recognition task.

The F1 descriptor has been proposed by Schindler and van Gool [24]. Both shape and motion flow features are extracted in a biologically-inspired fashion, by exploiting two parallel processing systems (see Figure 3) which bear similarity with the ventral and dorsal pathways of the visual cortex [23, 41]. The features are computed on a person-centered bounding box. Contrarily to silhouettes, bounding boxes can be reliably obtained in most scenarios, by using person detectors such as [42] or trackers based on rectangular windows [43]. By using this approach, we generate a single, 1000-dimensional feature vector for each frame of the sequence.

The F2 descriptor which we use has been proposed by Danafar and Gheissari in [19]. Such approach is based on histograms of optical flow [44], and captures both the local and global information of actions. The Harris corner detector [45] is first applied to extract interest points in each image; using the coordinates of the extracted interest points, bounding boxes are created around actors, and are vertically partitioned in three regions approximately corresponding to the head, torso, and legs (see Figure 4). The optical flow in each region is then computed by means of [44], and its horizontal and vertical components are quantized and represented as histograms. The resulting motion descriptors

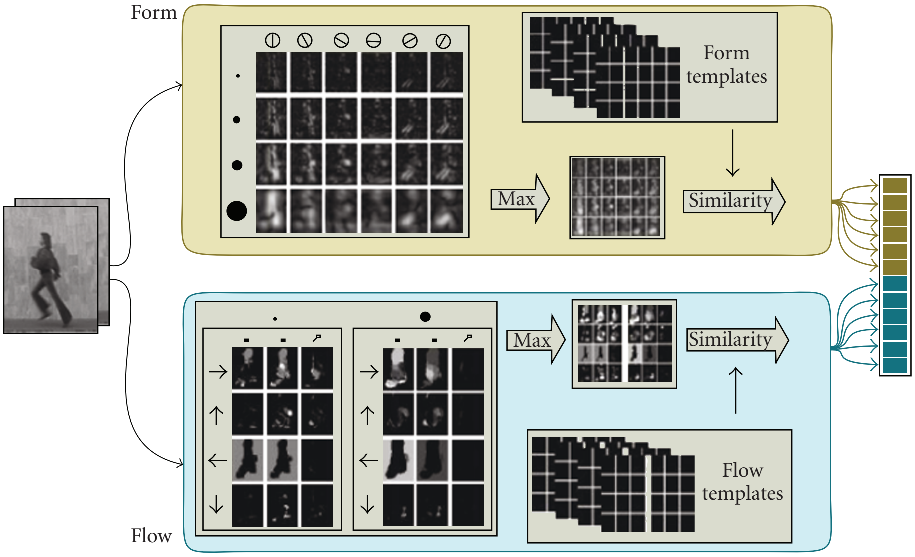


Figure 3: The features we use for unsupervised action recognition are described in [24], and are computed for each single frame of the video (left), while also considering the preceding one for computing optical flow. The resulting feature vector (right) consolidates data from two separate pathways computing form and flow, respectively. The former (top) computes gabor filter responses at different directions and scales; the letter in different directions, scales, and velocities. In both pathways, data is max-pooled for improving shift-invariance and summarized by matching with a set of templ

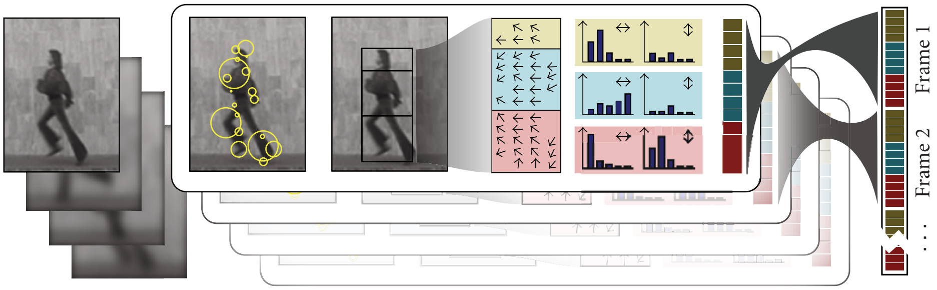


Figure 4: The features used in supervised case are described in [19]; a single feature vector (right) is computed for each sequence, by concatenating data coming from each frame of the sequence (left). In each frame, Harris interest points are used to recover a tight bounding box, which is vertically partitioned in three regions. The topmost 1/5 of the bounding box approximately contains the head and the neck. The middle 2/5 contains the torso and hands, whereas the bottom 2/5 of the bounding box contains the legs. Such segmentation is obviously approximated, and the resulting features would still be usable in cases where the assumptions are not met. Flow data in each region is summarized in separate histograms for the horizontal and vertical directions.

are computed from a combination of motion histograms for each of the three parts, originating a 102-dimensional feature vector for each frame. When using this approach, we finally combine the feature vectors for all frames in a single descriptor for the whole sequence.

7. Experiments and Evaluation

In order to gather experimental evidence that supports our proposed approach, we used two public datasets frequently referenced in the action recognition literature: the KTH human action database [1, 2] and the Weismann human action dataset [3].

The KTH dataset contains 2391 sequences of 6 types of human actions: walking, jogging, running, boxing, hand waving, and hand clapping. These actions were performed by 25 people in four different scenarios: outdoors (s1), outdoors with scale variations (s2), outdoors with different clothes (s3), and indoors (s4). Some samples from this dataset are shown in Figure 1.

The Weismann dataset contains 10 categories of actions: in accordance to several previous works [3, 24] we disregard the skip action and kept 9 distinctive categories: walk, run, jump, gallop sideways, bend, one-hand wave, two-hands wave, jump in place, and jumping jack. Each action was performed by 9 subjects. Example images from video sequences of this dataset are shown in Figure 6.

7.1. Unsupervised Classification. We tested unsupervised classification on both databases using feature F1. Mirroring

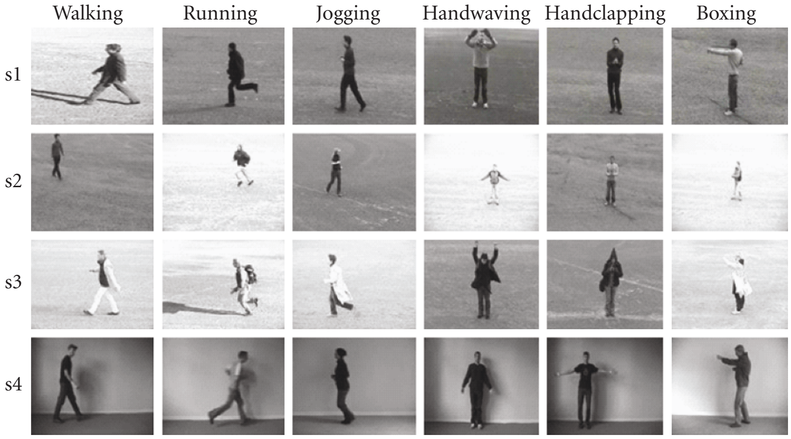


Figure 5: Example images from video sequences in KTH dataset (publicly available at http://www.nada.kth.se/cvap/actions/). This dataset was benchmarked with both unsupervised and supervised methods in the current study.

Figure 6: Example images from video sequences in Weismann dataset. We benchmarked on this dataset with the unsupervised method in the current study.

the experimental validation in [12], we considered a 27-frame sequence for each of the Weizmann videos, and a 17-frame sequence for KTH videos. We used the same bounding box data as in [12]. In particular, in the Weizmann dataset the fixed-size bounding boxes can be trivially extracted by considering a simple background subtraction algorithm. In the more challenging KTH dataset, bounding boxes for all frames are linearly interpolated from known initial and final positions.

Because of the small number of actors in Weizmann dataset, we evaluate the results with leave-one-out cross-validation. First, 72 unlabeled sequences from 8 subjects are used for recovering the 9 clusters in an unsupervised way; then, the 9 sequences from the one remaining subject are used for testing generalization capability. The procedure repeated for all 9 permutations. In the larger KTH dataset, we used a single partition of 16 subjects for clustering and 9 for testing generalization capability.

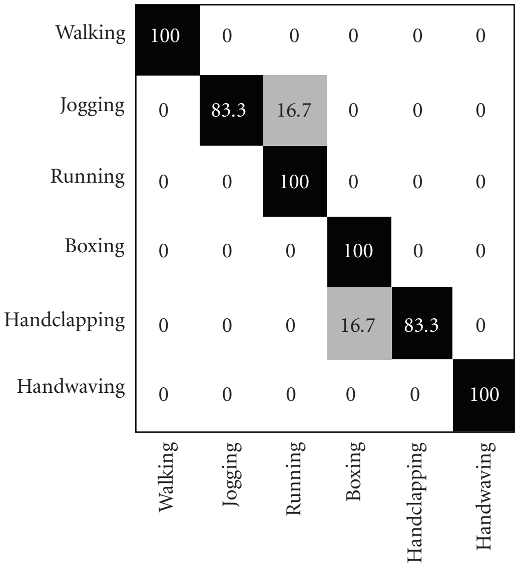
We report 100% accuracy for unsupervised classification in the Weizmann dataset, meaning that, in all the cross-validation folds, the 9 detected clusters actually coincide with the 9 different actions, and all the testing sequences are classified correctly. As reported in Table 1, this is the first time that perfect accuracy in unsupervised action recognition is reported on the Weizmann dataset—although other supervised techniques also reach 100% accuracy.

On the more complex KTH dataset, we obtained 94.4% overall recognition rate, which outperforms other reported methods, either supervised or unsupervised (see Table 2). Figure 6 reports the confusion matrix with our approach.

On the larger KTH dataset, training and testing took, respectively, 287 and 100 seconds, on a mid-level dual core laptop. The computational complexity is quadratic with respect to the number of frames in each sequence [8, 9]. The overall acceptance rate of $\mathcal{H}_0$, representing the similarity of two sequences, was computed from 100 runs of each homogeneity test.

7.2. Supervised Classification. For supervised classification, we worked with feature descriptor F2. On the KTH dataset, we considered subsequences of at most 150 frames, which are all summarized in a single feature vector. Such feature has proven to be less powerful than F1, which causes in the KTH dataset a recognition rate in the supervised case of 93.1%; this is lower than the 94.4% rate we obtained in the unsupervised case, when using the F1 features.

It is interesting to compare the effect of characteristic kernels, which we are using in this paper, to histogram intersection kernels, which are not characteristic and are widely used in the computer vision literature [37–39] for classifying histogram-based data. In fact, as reported in Section 4, characteristic kernels bear important advantages from the



	Walking	Jogging	Running	Boxing	Handclapping	Handwaving
Walking	100	0	0	0	0	0
Jogging	0	83.3	16.7	0	0	0
Running	0	0	100	0	0	0
Boxing	0	0	0	100	0	0
Handclapping	0	0	0	16.7	83.3	0
Handwaving	0	0	0	0	0	100

Figure 7: Confusion matrix achieved by our Unsupervised classification algorithm with MMD on KTH human action dataset. The overall accuracy rate of 94.4% is achieved with this method.

Table 1: Comparison of recognition results on Weismann dataset with different approaches.

Method	classification	Recognition rate %
MMD	Unsupervised	100
Schindler and Van Gool [24]	Supervised	100
Blank et al. [3]	Supervised	100
Niebles et al. [11]	Unsupervised	95
Jhuang et al. [20]	Supervised	98.8
Wang and Suter [15]	Supervised	97.8
Dollár et al. [21]	Supervised	86.7

theoretical point of view. Our results confirm such advantage in this practical application. Our reported accuracy of 93.1%, obtained with characteristic kernels, is a very significant improvement with respect to the accuracy of 85.3% reported in [19], obtained using histogram intersection kernels in the same setting.

We also compare our novel characteristic kernel for histogram-based data to the Gaussian kernel, which is also characteristic but is not tailored to histogram-based data. In our experiments, the accuracy of the Gaussian kernel is 33.8 %, which is much lower than our result of 93.1%. Confusion matrices in the three cases are reported in Figure 7.

Therefore, we can conclude that our experimental results are due to our kernel being both characteristic and suitable for histogram-based data, removing any of the two properties results in a significant performance loss.

Training and testing on the KTH dataset required 8 and 2 seconds, respectively. During the testing phase, the complexity is linear with the number of support vectors. The complexity of the training phase is dominated by the solution of a quadratic optimization problem.

Table 2: Comparison of recognition results on KTH dataset with different approaches. Note that the recognition rate reported by Jhuang et al. [20] is obtained on video sequences from scenarios 1 and 4 only. Other reported rates are on all scenarios.

Method	Classification	Recognition rate %
MMD	Unsupervised	94.4
SVM by charac. Kernel	Supervised	93.1
Schindler and Van Gool [24]	Supervised	92.7
Jhuang et al. [20]	Supervised	91.7
Nowozin et al. [31]	Supervised	87
Wong and Cipolla [32]	Supervised	86.6
Danafar and Gheissari [19]	Supervised	85.3
Niebles et al. [11]	Unsupervised	83.3
Dollár et al. [21]	Supervised	81.2
Schüldt et al. [2]	Supervised	71.7

Given training vectors $x_i \in \mathbb{R}^n$, $i = 1, \dots, l$, in two classes, and a vector $y \in R^l$, C-SVC solves a quadratic problem to find support vectors which formulated as:

$$\begin{aligned} &\min_\alpha \quad \frac{1}{2}\alpha^T Q\alpha - e^T\alpha, \\ &\text{subject to} \quad y^T\alpha = 0, \quad 0 \le \alpha_i < C, \; i = 1, \dots, l, \end{aligned} \tag{17}$$

where e is the vector of all ones, $C > 0$ (we tuned $C = 1$) is the upper bound, Q is a $l \times l$ positive semidefinite matrix

$$Q_{ij} \equiv y_i y_j k\left(x_i, x_j\right), \qquad k\left(x_i, x_j\right) \equiv \phi(x_i)^T \phi\left(x_j\right) \tag{18}$$

is the kernel. The difficulty of solving the above equation is the density of Q, whose elements are in general not zero. To overcome this problem the decomposition method is implemented. The time complexity is at most $O(nl^2)$ if we suppose each kernel evaluation is $O(n)$ [46]. The time performance for training and testing are, respectively 8, and 2 seconds.

In our case we deal with multiclass type of classification, and we consider one-vs-one procedure. Thus, if m is the number of classes (actions), $m(m-1)/2$ comparisons are needed (in our case $m = 6$).

8. Conclusions

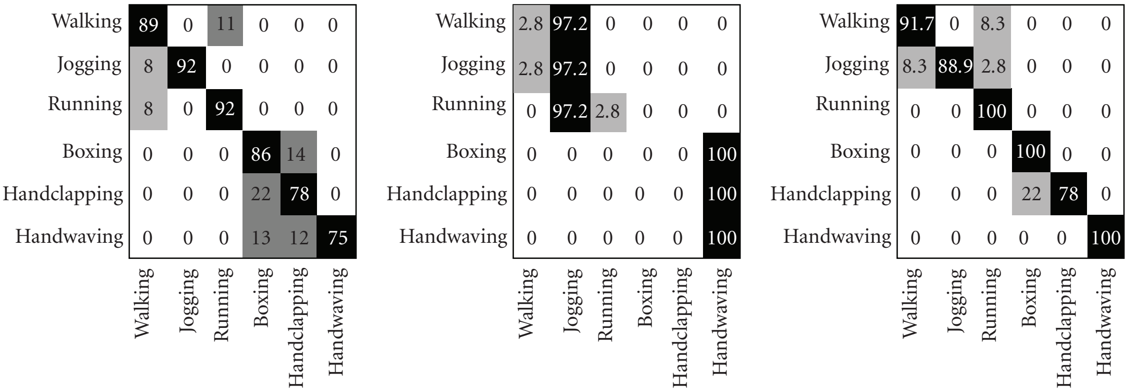
We successfully dealt with the challenging task of recognizing actions in videos. In particular, we described

(i) an unsupervised nonparametric kernel method based on Maximum Mean Discrepancy,

(ii) a supervised method using Support Vector Machines with a novel proper characteristic kernel for Abelian semigroups.

On the two major data sets for action recognition, our approaches outperformed those found in the literature, both in the unsupervised and supervised case.

The new characteristic kernel is suitable for histograms, and may be useful for many other computer vision problems involving histogram-based features.



	Walking	Jogging	Running	Boxing	Handclapping	Handwaving
Walking	89	0	11	0	0	0
Jogging	8	92	0	0	0	0
Running	8	0	92	0	0	0
Boxing	0	0	0	86	14	0
Handclapping	0	0	0	22	78	0
Handwaving	0	0	0	13	12	75

(a) Results with histogram intersection kernel as a positive definite kernel with overall accuracy rate of 85.3%

	Walking	Jogging	Running	Boxing	Handclapping	Handwaving
Walking	2.8	97.2	0	0	0	0
Jogging	2.8	97.2	0	0	0	0
Running	0	97.2	2.8	0	0	0
Boxing	0	0	0	0	0	100
Handclapping	0	0	0	0	0	100
Handwaving	0	0	0	0	0	100

(b) Results with Gaussian kernel as a characteristic kernel with overall accuracy rate of 33.8%

	Walking	Jogging	Running	Boxing	Handclapping	Handwaving
Walking	91.7	0	8.3	0	0	0
Jogging	8.3	88.9	2.8	0	0	0
Running	0	0	100	0	0	0
Boxing	0	0	0	100	0	0
Handclapping	0	0	0	22	78	0
Handwaving	0	0	0	0	0	100

(c) Results with our proposed kernel k_{HC}, with overall accuracy rate of 93.1%

FIGURE 8: Confusion matrices obtained on the KTH dataset with F2 descriptors [19], using SVM and the indicated kernels; (a) shows the rec with recognition rates of histogram Intersection kernel which is a positive definite but not a characteristic kernel; (b) denotes the result of a characteristic kernel (Gaussian) which is not tailored for histogram based information; (c) is the result of characteristic kernel which is tailored for histograms.

Acknowledgment

This work was partially funded by SNF Sinergia grant number CRSIKO_122697/1.

References

[1] I. Laptev and T. Lindeberg, "Local descriptors for spatio-temporal recognition," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, 2003.

[2] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local SVM approach," in *Proceedings of the 17th International Conference on Pattern Recognition (ICPR '04)*, pp. 32–36, August 2004.

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, pp. 1395–1402, October 2005.

[4] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reporoducing kernel Hilbert spaces," *Journal of Machine Learning Research*, vol. 5, pp. 73–99, 2004.

[5] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, "Kernel measures of conditional dependence," *Advances in Neural Information Processing Systems*, vol. 20, pp. 489–496, 2008.

[6] K. Fukumizu, B. K. Sriperumbudur, A. Gretton, and B. Schölkopf, "Characteristic kernels on groups and semigroups," in *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems(NIPS '08)*, 2008.

[7] B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckreit, and B. Schölkopf, "Injective Hilbert space embeddings of probability measures," in *Proceedings of the 21st Annual Conference on Learning Theory (COLT '08)*, R. Servedio and T. Zhang, Eds., pp. 111–122, Springer, July 2008.

[8] A. Gretton, K. Borgwardt, M. Rasch, A. Smola, and B. Schölkopf, "A kenel method for the two-sample problem," in *Proceedings of the 19th Conference on Advances in Neural Information Processing Systems*, B. Schölkopf, J. Platt, and T. Hoffman, Eds., pp. 513–520, MIT Press, Vancouver, Canada, 2006.

[9] A. Gretton, K. Borgwardt, M. Rasch, A. Smola, and B. Schölkopf, "A kenel method for the two-sample problem," Tech. Rep. 157, Max-Planck-Institut for Biological Cybernetics, 2008.

[10] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatio-temporal words," in *Proceedings of the British Machine Vision Conference (BMVC '06)*, 2006.

[11] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International Journal of Computer Vision*, vol. 79, no. 3, pp. 299–318, 2008.

[12] P. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea, "Machine recognition of human activities: a survey," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1473–1488, 2008.

[13] A. F. Bobick and J. W. Davis, "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[14] S. Carlson and J. Sullivan, "Action recognition by shape matching to key frames," in *Proceedings of the Workshop on Models versus Exemplars in Computer Vision*, 2001.

[15] L. Wang and D. Suter, "Recognizing human activities from silhouettes: motion subspace and factorial discriminative graphical model," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '07)*, pp. 1–8, Minneapolis, Minn, USA, June 2007.

[16] A. Yilmaz and M. Shah, "Actions sketch: a novel action representation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, pp. 984–989, IEEE Computer Society, June 2005.

[17] A. A. Efros, A. C. Berg, G. Mori, and J. Malik, "Recognizing action at a distance," in *Proceedings of the 9th IEEE International Conference on Computer Vision (ICCV '03)*, pp. 726–733, October 2003.

[18] E. Shechtman and M. Irani, “Space-time behavior based correlation,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’05)*, pp. 405–412, June 2005.

[19] S. Danafar and N. Gheissari, “Action recognition for surveillance application using optic flow and SVM,” in *Proceedings of the 8th Asian Conference on Computer Vision (ACCV ’07)*, Tokyo, Japan, November 2007.

[20] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, “A biologically inspired system for action recognition,” in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV ’07)*, Rio de Janeiro, Brazil, October 2007.

[21] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, “Behavior recognition via sparse spatio-temporal features,” in *Proceedings of the 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS ’05)*, pp. 65–72, October 2005.

[22] Y. Ke, R. Sukthankar, and M. Hebert, “Spatio-temporal shape and flow correlation for action recognition,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’07)*, pp. 1–8, June 2007.

[23] M. A. Giese and T. Poggio, “Neural mechanisms for the recognition of biological movements,” *Nature Reviews Neuroscience*, vol. 4, no. 3, pp. 179–192, 2003.

[24] K. Schindler and L. Van Gool, “Action snippets: how many frames does human action recognition require?” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR ’08)*, pp. 1–8, June 2008.

[25] A. Gilbert, J. Illingworth, and R. Bowden, “Scale invariant action recognition using compound features mined from dense spatio-temporal corners,” in *Proceedings of the 10th European Conference on Computer Vision (ECCV ’08)*, pp. 222–233, Marseille, France, 2008.

[26] J. Hoey, “Hierarchical unsupervised learning of facial expression categories,” in *Proceedings of the IEEE Workshop on Detection and Recognition of Action Video*, pp. 99–106, 2001.

[27] H. Zhong, J. Shi, and M. Visontai, “Detecting unusual activity in video,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’04)*, pp. 819–826, July 2004.

[28] T. Xiang and S. Gong, “Video behaviour profiling and abnormality detection without manual labelling,” in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV ’05)*, pp. 1238–1245, Beijing, China, October 2005.

[29] O. Boiman and M. Irani, “Detecting irregularities in images and in video,” in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV ’05)*, vol. 1, pp. 462–469, IEEE Computer Society, October 2005.

[30] Y. Wang, H. Jiang, M. S. Drew, Z.-N. Li, and G. Mori, “Unsupervised discovery of action classes,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR ’06)*, vol. 2, pp. 1654–1661, June 2006.

[31] S. Nowozin, G. Bakir, and K. Tsuda, “Discriminative subsequence mining for action classification,” in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV ’07)*, Rio de Janeiro, Brazil, October 2007.

[32] S.-F. Wong and R. Cipolla, “Extracting spatiotemporal interest points using global information,” in *Proceedings of the IEEE 11th International Conference on Computer Vision (ICCV ’07)*, pp. 1–8, Rio de Janeiro, Brazil, October 2007.

[33] J. Pei, J. Han, B. Mortazavi-Asl, et al., “Mining sequential patterns by pattern-growth: the prefixspan approach,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 11, pp. 1424–1440, 2004.

[34] A. Demiriz, K. P. Bennett, and J. Shawe-Taylor, “Linear programming boosting via column generation,” *Journal of Machine Learning*, vol. 46, no. 1–3, pp. 225–254, 2002.

[35] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola, “Integrating structured biological data by kernel maximum mean discrepancy,” *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[36] Z. Hachaoui, F. Bach, and E. Moulines, “Testing for homogeneity with kernel fisher discriminant analysis,” in *Proceedings of the 22nd Annual Conference on Neural Information Processing Systems (NIPS ’08)*, vol. 20, pp. 609–616, Vancouver, Canada, December 2008.

[37] M. J. Swain and D. H. Ballard, “Color indexing,” *International Journal of Computer Vision*, vol. 7, no. 1, pp. 11–32, 1991.

[38] F. Odone, A. Barla, and A. Verri, “Building kernels from binary strings for image matching,” *IEEE Transactions on Image Processing*, vol. 14, no. 2, pp. 169–180, 2005.

[39] S. Boughorbel, J.-P. Tarel, and N. Boujemaa, “Generalized histogram intersection kernel for image recognition,” in *Proceedings of the IEEE International Conference on Image Processing (ICIP ’05)*, pp. 161–164, Genoa, Italy, September 2005.

[40] K. Fukushima, “Neocognitron: a self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position,” *Biological Cybernetics*, vol. 36, no. 4, pp. 193–202, 1980.

[41] D. J. Field, “Relations between the statistics of natural images and the response properties of cortical cells,” *Journal of the Optical Society of America A*, vol. 4, no. 12, pp. 2379–2394, 1987.

[42] A. Casile and M. A. Giese, “Critical features for the recognition of biological motion,” *Journal of Vision*, vol. 5, no. 4, pp. 348–360, 2005.

[43] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV ’05)*, pp. 886–893, October 2005.

[44] M. J. Black and P. Anandan, “The robust estimation of multiple motions: parametric and piecewise-smooth flow fields,” *Computer Vision and Image Understanding*, vol. 63, no. 1, pp. 75–104, 1996.

[45] C. Harris and M. Stephens, “A combined corner and edge detector,” in *Proceedings of the 4th Alvey Vision Conference*, pp. 147–151, Manchester, UK, 1988.

[46] C.-C. Chang and C.-J. Lin, “LIBSVM: a library for Support Vector Machines,” 2009, http://www.csie.ntu.edu.tw/~cjlin/libsvm/.