Multimed Tools Appl (2010) 48:69–87
DOI 10.1007/s11042-009-0351-3

Video event classification using string kernels

Lamberto Ballan · Marco Bertini · Alberto Del Bimbo · Giuseppe Serra

Published online: 15 September 2009

Abstract Event recognition is a crucial task to provide high-level semantic description of the video content. The bag-of-words (BoW) approach has proven to be successful for the categorization of objects and scenes in images, but it is unable to model temporal information between consecutive frames. In this paper we present a method to introduce temporal information for video event recognition within the BoW approach. Events are modeled as a sequence composed of histograms of visual features, computed from each frame using the traditional BoW. The sequences are treated as strings (*phrases*) where each histogram is considered as a character. Event classification of these sequences of variable length, depending on the duration of the video clips, are performed using SVM classifiers with a string kernel that uses the Needlemann-Wunsch edit distance. Experimental results, performed on two domains, soccer videos and a subset of TRECVID 2005 news videos, demonstrate the validity of the proposed approach.

Keywords Video annotation · Event classification · Bag-of-words · String kernel · Edit distance

L. Ballan (✉) · M. Bertini · A. Del Bimbo · G. Serra
Media Integration and Communication Center, University of Florence, Florence, Italy
e-mail: ballan@dsi.unifi.it

M. Bertini
e-mail: bertini@dsi.unifi.it

A. Del Bimbo
e-mail: delbimbo@dsi.unifi.it

G. Serra
e-mail: serra@dsi.unifi.it

1 Introduction

Recently it has been shown that part-based approaches are effective methods for object detection and recognition due to the fact that they can cope with partial occlusions, clutter and geometrical transformations. Many approaches have been presented, but a common idea is to model a complex object or a scene by a collection of local interest points. Each of these local features describes a small region around the interest point and therefore they are robust against occlusion and clutter. To achieve robustness to changes of viewing conditions the features should be invariant to geometrical transformations such as translation, rotation, scaling and also affine transformations. In particular, SIFT features by Lowe [26] have become the de facto standard because of their high performances and (relatively) low computational cost. In fact, SIFT features have been frequently applied to object or scene recognition and also to many other related tasks. In this field, a solution that recently has become very popular is the Bag-of-Words (BoW) approach. It has been originally proposed for natural language processing and information retrieval, where it is used for document categorization in a text corpus, where each document is represented by its word frequency. In the visual domain, an image or a frame of a video is the visual analogue of a document and it can be represented by a bag of quantized invariant local descriptors (usually SIFT), called *visual-words* or *visterms*. The main reason for its success is that it provides methods that are sufficiently generic to cope with many object types simultaneously. We are thus confronted with the problem of generic visual categorization [14, 38, 44, 46], like classification of objects or scenes, instead of recognizing a specific class of objects. The efficacy of the BoW approach is demonstrated also by the large number of systems based on this approach that participate in the PASCAL VOC and TRECVID [39] challenges.

More recently, part-based models have been successfully applied also to the classification of human actions [12, 33], typically using salient features that represent also temporal information (such as spatio-temporal gradients), and to video event recognition. These tasks are particularly important for video indexing and retrieval where dynamic concepts occur very frequently. Even if few novel spatio-temporal features have been proposed, the most common solution is to apply the traditional BoW approach using static features (e.g. SIFT) on a keyframe basis. Unfortunately, for this purpose the standard BoW approach has shown some drawbacks with respect to the traditional image categorization task. Perhaps the most evident problem is that it does not take into account temporal relations between consecutive frames. In this way, event detection suffers from the incomplete dynamic representation given by the keyframe, resulting in a poor performance compared to the results obtained for the detection of static concepts. Figure 1 shows a few examples of difficulties that arise when performing event detection using only keyframes. Nevertheless, only few works have been proposed to cope with this problem [40, 42].

In this paper, we present a novel method to model actions as a sequence of histograms (one for each frame) represented by a traditional bag-of-words approach. An action is described by a "phrase" of variable length, depending on the clip's duration, thus providing a global description of the video content that is able to incorporate temporal relations. Then video phrases can be compared by computing edit distances between them and, in particular, we use the Needleman-Wunsch distance [31] because it performs a global alignment on sequences dealing with

(a) (b) (c)

Fig. 1 Keyframe-based video event detection. **a** Is it *shot-on-goal* or *placed-kick*? **b** Is it *walking* or *running*? **c** Is it a *car exiting* or *entering* from somewhere?

video clips of different lengths. Using this kind of representation we are able to perform classification of video events and, following the promising results obtained in text categorization [25] and in bioinformatics (e.g. protein classification) [24], we investigate the use of SVMs with a string kernel, based on edit distance, to perform classification. Experiments have been performed on soccer and news video datasets, comparing the proposed method to a baseline kNN classifier and to a traditional keyframe-based BoW approach. Experimental results obtained by SVM and string kernels outperform the other approaches and, more generally, they demonstrate the validity of the proposed method.

The rest of the paper is organized as follows. A review of related previous works is presented in the next section. The techniques for frame and event representation are discussed in Section 3. The classification method, including details about the SVM string kernel, is presented in Section 4. Experimental results are discussed in Section 5 and, finally, conclusions are drawn in Section 6. A preliminary version of this paper, focused only on the soccer domain, appeared in CBMI 2009 [2].

2 Related works

Video event detection and recognition is really challenging because of complex motion, occlusions, clutter, geometric transformations and illumination changes. Nevertheless, it is an essential task for automatic video content analysis and annotation. Previous works in this field can be roughly grouped into three main categories; abnormal/unusual event detection [7, 41, 45], human action categorization [6, 12, 23, 33, 36], and video event recognition [13, 19, 20, 40, 42].

Unusual event detection and activity recognition are very active research areas in video surveillance and many different approaches have been previously proposed. Several of these works rely on HMM models or Dynamic Bayesian Networks. In [45], Zhang et al. used HMMs to model usual events from a large training set; unusual event models are learned in a second step through Bayesian adaptation. The problem of detecting suspicious behaviors in video sequences is addressed also by Boiman and Irani [7]. They posed the problem as an inference process in a probabilistic graphical model, used to describe large ensembles of patches at multiple spatio-temporal scales. Inferred unusual configurations are treated as suspicious behaviors.

Over the past decade, the specific problem of recognizing human actions has received considerable attention from the research community. In fact, an automatic human activity recognition method may be very useful for many applications such as video surveillance, video annotation and retrieval and human–computer interaction. The early works in this field are usually based on holistic representations. For example, Bobick et al. [6] proposed motion history images to encode short spans of motion; this representation is then matched using global statistics, such as moment features. Although this method is efficient, it is assumed to have a well segmented foreground and background. More recently, part-based appearance models have been successfully applied to the human action categorization problem, because they overcome some limitations of holistic models such as the necessity of performing background subtraction and tracking. These approaches rely on salient visual features that represent also temporal information (such as spatio-temporal intensity gradients) or motion descriptors like optical flow. Laptev [22] proposed a spatio-temporal interest point detector by extending the Harris corner operator. Local features are extracted from locations of the video which exhibit strong variations of intensity both in spatial and temporal directions. Dollar et al. [12] applied separable linear Gabor filters, treating time differently from space and looking for locally periodic motion. These features have been frequently used by different researchers within part-based frameworks (e.g. the BoW approach) in combination to learning techniques such as support vector machines (SVM) [36] and probabilistic latent semantic analysis (pLSA) [33]. More recently, Laptev et al. [23] have abandoned the interest point detection approach, preferring a structural representation based on dense temporal and spatial scale sampling (inspired by spatial pyramids), providing state-of-the-art results and showing promising results also on realistic video settings. Instead, in [20] action is modeled by a space-time volume in the video sequence and volumetric features (based on optic flow) are extracted for event detection. However, the performance of these methods heavily depends on the spatio-temporal features which often privilege high-motion regions. As a result, the approach is very sensitive to motion, thus providing high performance in the recognition of motion events that are more frequent in constrained video domains such as videosurveillance.

The generalization of this approach to less constrained and more general domains, like news videos or movies, has not been demonstrated. Therefore, the most common solution is to apply the traditional BoW approach using static features (such as SIFT descriptors) on a keyframe basis; in fact, many of the methods that have been submitted to the TRECVID competition extend this idea. Unfortunately, the application of this approach to event classification has shown some drawbacks with respect to the traditional image categorization task. The main problem is that it does not take into account temporal relations between consecutive frames, and thus event classification suffers from the incomplete dynamic representation. Recently some attempts have been made to employ temporal information among static part-based representations of video frames. Xu and Chang [42] proposed to apply Earth Mover's Distance (EMD) and Temporally Aligned Pyramid Matching (TAPM) for measuring video similarity; EMD distance is incorporated in a SVM framework for event detection in news videos. In [40], BoW is extended constructing relative motion histograms between visual words (ERMH-BoW) in order to employ motion relativity and visual relatedness. Zhou et al. [47] presented a SIFT-Bag based generative-to-discriminative framework for video event detection, providing

improvements on the best results of [42] on the same TRECVID 2005 corpus. They proposed to describe video clips as a bag of SIFT descriptors by modeling their distribution with a Gaussian Mixture Model (GMM); in the discriminative stage, specialized GMMs are built for each clip and video event classification is performed.

Similar approaches for event detection in news videos have been applied also at a higher semantic level, using the scores provided by concept detectors as synthetic frame representations or exploiting some pre-defined relationships between concepts. For example, Ebadollahi et al. [13] proposed to treat each frame in a video as an observation, applying then HMM to model the temporal evolution of an event. Yang and Hauptmann [43] proposed to exploit temporal consistency between nearby shots (described by their concept score) obtaining a temporal smoothing procedure for improving video retrieval. In [16] an ontology framework (VERL) has been defined for representation and annotation of video events. Finally, Bertini et al. [5] have recently presented an ontology-based framework for semantic video annotation by learning spatio-temporal rules; in their approach, an adaptation of the First Order Inductive Learner (FOIL) is used to learn rule patterns that have been then validated on few TRECVID 2005 video events.

3 Event representation and classification

Given a set of labeled videos, our goal is to automatically learn event models to perform categorization of new videos. In this work, we investigate in particular a new way of representing an event and how to learn this representation. An overview of our approach is illustrated in Fig. 2.

Structurally an event is represented by a sequence of frames, that may have different lengths depending on how it has been carried out. We model an event by a sequence of visual word frequency vectors, computed from the frames of the sequence; considering each frequency vector as a *character* we call this sequence (i.e. string) *phrase*. Additionally, we define a kernel, based on an edit-distance, used by SVMs to handle variable-length input data such as this kind of event representation.

3.1 Frame representation

Video frames are represented using bag-of-words, because this representation has demonstrated to be flexible and effective for various image analysis tasks [14, 38, 46]. First of all, a visual vocabulary is obtained through vector quantization of large sets of local feature descriptors extracted from a collection of videos. In this work, we use DoG [28] as keypoint detector and SIFT [26] as keypoint descriptor. The visual vocabulary (or codebook) is generated by clustering the detected keypoints in the feature space using the k-means algorithm and Euclidean distance as the clustering metric. The center of each resulting cluster is defined as *visual word*. The size of the visual vocabulary is determined by the number of clusters and it is one of the main critical point of the approach. A small vocabulary may lack discriminative power since two features may be assigned to the same cluster even if they are not similar, while a large vocabulary is less generalizable. The trade-off between discrimination and generalization is highly content dependent and it is usually determined by

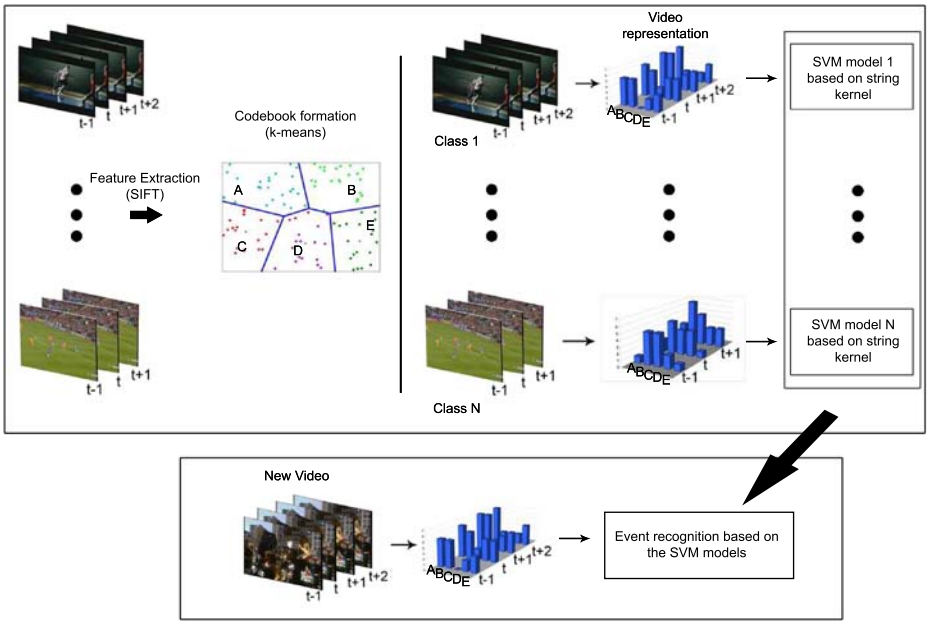


Fig. 2 Schematization of the proposed approach. In the training stage the features (SIFT) extracted from videos are clustered into visual words (A,B,C,D,E). Each video is represented as a sequence of BoW histograms. Events are described by a *phrase* (string) of variable length, depending on the clip's duration. SVMs with string kernel are used to learn the event representation model for each class. The learned models can be used to recognize events in a new video

experiments [44]. The effect of the codebook size is explored also in our experiments (see Section 5). Once a vocabulary is defined, each detected keypoint in a frame is assigned to a unique cluster membership (i.e. a particular visual word), so that a frame is represented by a visual word frequency vector. In this way, this frame representation ignores the spatial arrangement between the different words and thus between the extracted visual features. This effect brings the advantages of using a simple representation that makes learning efficient but, on the other hand, it discards useful information. Alternative approaches might include structural information by encoding information of the structure of the model, for example by modeling the geometrical arrangement of local features [15]. In most cases, the trade-off is an increased computational complexity.

3.2 Video representation

As previously introduced, each video shot is described as a *phrase* (string) formed by the concatenation of the bag-of-words representations of consecutive *characters* (frames). To compare these *phrases*, and consequently actions and events, we can adapt metrics defined in the information theory.

Edit distance The edit distance between two string of characters is the number of operations required to transform one of them into the other. There are several different algorithms to define or calculate this metric, and different transformations

can be used. In particular, our approach uses the Needleman-Wunsch distance [31] with the substitution, insertion and deletion transformations. The main motivation of this choice is that Needleman-Wunsch distance performs a global alignment that accounts for the structure of the strings, and the distance can be considered as a score of similarity. The basic idea of the algorithm is to build up the best alignment through optimal alignments of smaller subsequences, using dynamic programming; unlike other approaches, such as dynamic time warping, this type of edit distance algorithms is able to cope with noise and inaccurate sequence segmentation [34]. Considering the cost matrix C that tracks the costs of the edit operations needed to match two strings, we can then write the cost formula for the alignment of the a_i and b_j characters of two strings as:

$$C_{i,j} = min\left(C_{i-1,j-1} + \delta\left(a_i, b_j\right), C_{i-1,j} + \delta_I, C_{i,j-1} + \delta_D\right)$$

where $\delta(a_i, b_j)$ is 0 if the distance between a_i and b_j is close enough to evaluate $a_i \approx b_j$ or the cost of substitution otherwise, δ_I and δ_D are the costs of insertion and deletion, respectively. The matrix contains all possible pair combinations that can be constructed from the two sequences being compared, and every possible comparison of the sequences can be represented by a path in the matrix. Figure 3 shows an example of the evaluation of the Needleman-Wunsch distance for the case of text and soccer action, respectively. The distance is the number in the lower-right corner of the cost matrix. The traceback that shows the sequence of edit operations leading to the best alignment between the sequences is highlighted in each cost matrix. The algorithm guarantees to find the best alignment of the sequences and is $O(mn)$ in time and space, where m and n are the lengths of the two strings being compared. We

(a) text example

		S	E	N	D
	0	1	2	3	4
A	1	1	2	3	4
N	2	2	2	2	3
D	3	3	3	3	2

(b) video example

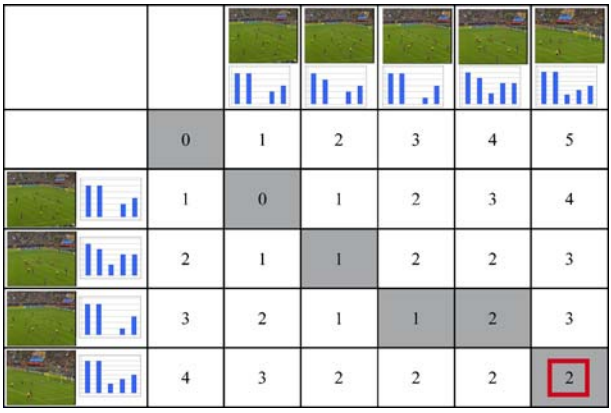


Fig. 3 Needleman-Wunsch edit distance: **a** text and **b** video examples. Each video frame is represented using its visual word frequency vector. The highlighted path in the cost matrix shows the sequence of operations leading to the best alignment

have used a dynamic programming implementation of the algorithm that reduces the space complexity to $O(min(m, n))$ [30].

Measuring similarity between characters A crucial point is the evaluation of the similarity among characters ($a_i \approx b_j$). In fact, when evaluating this similarity on text it is possible to define a similarity matrix between characters, because their number is limited. Instead, in our case each frequency vectors is a different character, therefore we deal with an extremely large alphabet. This requires us to define a function that evaluates the similarity of two characters. Since in our approach each character is an histogram we have evaluated several different methods to compare the frequency vectors of two frames, p and q. In particular we have considered the following distances: *Chi-square* test, *Kolmogorov-Smirnov* test, *Bhattacharyya*, *Intersection*, *Correlation*, *Mahalanobis*. Note that in our implementation each histogram is normalized to sum to one so that it can be considered as a probability distribution.

Chi-square test is a statistical method that permits to compare an observed frequency with a reference frequency. It is defined as:

$$d(p, q) = \sum_{k=1}^{N} \frac{(p(k) - q(k))^2}{p(k) + q(k)}. \tag{1}$$

A low value means a better match than a high value.

Kolmogorov-Smirnov test is a statistical method that quantifies the distance between one cumulative distribution function and a reference cumulative distribution function. In our case it can be defined as:

$$d(p, q) = \sup_k \left| F_p(k) - F_q(k) \right|, \tag{2}$$

where $F_s(k) = \sum_{j=1}^{k} s(j)$.

Bhattacharyya's distance is defined equal to:

$$d(p, q) = \left(1 - \sum_{k=1}^{N} \frac{\sqrt{p(k)q(k)}}{\sqrt{\sum_{k=1}^{N} p(k) \cdot \sum_{k=1}^{N} q(k)}} \right)^{\frac{1}{2}}. \tag{3}$$

Using this distance a perfect match is evaluated as 0, whereas a total mismatch is 1.

Intersection distance is equal to:

$$d(p, q) = \sum_{k=1}^{N} min(p(k), q(k)). \tag{4}$$

The intersection of two histograms is connected to the Bayes error rate, the minimum misclassification (or error) probability which is computed as the overlap between two PDF's $P(A)$ and $P(B)$. If both histograms are normalized to 1, then a perfect match is 1 and a total mismatch is 0.

Correlation is defined as:

$$d(p,q) = \frac{\sum_{k=1}^{N} p'(k)q'(k)}{\sqrt{\sum_{k=1}^{N} p'^2(k)q'^2(k)}}, \tag{5}$$

where $s'(k) = s(k) - (1/N)(\sum_{j=1}^{N} s(j))$ and N equals the number of bins in the histogram. For correlation, a high score represents a better match than a low score.

Mahalanobis is a distance between an unknown sample and a set of samples which has known mean vector and covariance matrix. Formally given a sample x and a group of samples Y with mean μ and covariance matrix Σ the Mahalanobis distance is:

$$d(x,Y) = (x-\mu)'\Sigma^{-1}(x-\mu). \tag{6}$$

In our case this distance can be exploited to find the similarity between a frequency vector of a frame p and a set of frames $q_{-n}.., q_{-1}, q, q_1, ..q_n$, where q_{-n} is n^{th} frame before q. In particular n is empirically set to ten.

4 Classification using string kernels

In recent years, Support Vector Machines (SVMs), introduced by Boser et al. [8], have become an extremely popular tool for solving classification problems. In their simplest version, given a set of labeled training vectors of two classes, SVMs map these vectors in a high dimensional space and learn a linear decision boundary between the two classes that maximizes the margin, which is defined to be the smallest distance between the decision boundary and any of the input samples. The result is a linear classifier that can be used to classify new input data. In the binary classification problem, suppose to have a training data set that comprises N input vectors $\mathbf{x}_1, ..., \mathbf{x}_N$, with corresponding target values $t_1, ..., t_N$ where $t_n \in \{-1, 1\}$. The SVMs approach finds the linear decision boundary $y(\mathbf{x})$ as:

$$y(\mathbf{x}) = w^T\phi(\mathbf{x}) + b \tag{7}$$

where ϕ denotes a fixed feature-space transformation, b is a bias parameter, so that, if the training data set is linearly separable, $y(\mathbf{x}_n) > 0$ for points having $t_n = +1$ and $y(\mathbf{x}_n) < 0$ for points having $t_n = -1$. In this case the maximum marginal solution is found by solving for the optimal weight vector $\mathbf{a} = (a_1, \ldots, a_N)$ in the dual problem in which we maximize:

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m t_n t_m \langle\phi(\mathbf{x}_n), \phi(\mathbf{x}_m)\rangle \tag{8}$$

with respect to $\mathbf{a}$, that is subject to the constraints:

$$\sum_{n=1}^{N} a_n t_n = 0, \quad a_n \geq 0 \quad \text{for} \quad n = 1, ..., N \tag{9}$$

where $\langle \phi(\mathbf{x}_n), \phi(\mathbf{x}_m) \rangle$ is the inner product of $\mathbf{x}_n$ and $\mathbf{x}_m$ in the feature-space. The parameters w and b are then derived from the optimal $\mathbf{a}$.

The mapping to a higher dimensionality feature-space is motivated by Cover's theorem [11]. This theorem states that a complex classification problem cast non-linearly into high dimensional space is more likely to be linearly separable than in the original low dimensional space. However, the explicit mapping of input samples in a high dimensional space, and then their inner product, generally have very high computational costs. Kernel functions have been introduced to handle this problem, since they permit to perform the inner product in the feature-space without requiring to explicitly perform the transformation $\langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle$. Formally, let χ be the original input vector space and $k : \chi \times \chi \rightarrow \Re$ a function mapping pairs of input vector to real numbers. If the function satisfies the Mercer condition [4] then there exists a feature-space and a mapping function ϕ such that k acts as a inner product in this feature-space and it is called valid kernel function [37]. In particular a necessary and sufficient condition for a function $k(\mathbf{x}_1, \mathbf{x}_2)$ to be a valid kernel is that the Gram matrix $\mathbf{K}$, whose elements are given by $k(\mathbf{x}_n, \mathbf{x}_m)$, should be positive semidefinite for all possible choices of the input samples.

Recently, many approaches in image categorization have successfully used different kernels such as linear, radial and chi-square basis functions; in particular the latter often gives the best results [46]. However, these kernels are not appropriate for event classification. In fact they deal with input vectors with fixed dimensionality, whereas vectors that represent an action usually have different lengths, depending on how the action is performed. Unlike other approaches that solve this problem simply by representing the video clips with a fixed number of samples [35], we introduce a kernel that deals with input vectors with different dimensionality, in order to account for the temporal progression of the actions. Starting from a Gaussian Kernel that takes the form:

$$k(\mathbf{x}, \mathbf{x}') = exp(-\left\|\mathbf{x} - \mathbf{x}'\right\|^2 / 2\sigma^2), \tag{10}$$

we replace the Euclidean with the Needleman-Wunsch (NW) [31] edit distance. Thus the proposed resulting kernel is:

$$k(\mathbf{x}, \mathbf{x}') = exp(-d(\mathbf{x}, \mathbf{x}')). \tag{11}$$

where $d(\mathbf{x}, \mathbf{x}')$ is the NW edit distance between $\mathbf{x}$ and $\mathbf{x}'$ input vectors.

It has been previously empirically demonstrated that this type of kernel obtains good results for classification of handwritten digits, shapes, chromosome images [1, 29, 32], despite the fact that the edit distance has not been proved to be a valid kernel. These good empirical results recently have become subject of investigation, in order to obtain a more formal theoretical understanding for the use of indefinite kernel functions [10, 18, 27]. In the cases in which the kernel does not satisfy the Mercer condition, it is possible to adjust the Gram matrix by adapting eigenvalues of this matrix to be all positive, as described in [17]. However, it should be noted that the Gram matrices we applied in our experiments did not require any adaptation.

Fig. 4 Soccer dataset consists of four different events: shot-on-goal, placed-kick, throw-in and goal-kick

5 Experimental results

We have carried out video event classification experiments on different domains, soccer videos and a subset of TRECVID 2005 video corpus, to analyze the performance of the proposed method and to evaluate its general applicability. The soccer dataset consists of 100 video clips in MPEG-2 format at full PAL resolution (720 × 576 pixels, 25 fps), and it contains 4 different events: *shot-on-goal*, *placed-kick*, *throw-in* and *goal-kick*. Examples of these events are shown in Fig. 4. The full dataset, including also ground truth, is available on request at our webpage.[1] The sequences were taken from 5 different matches of the Italian *"Serie A"* league (season 2007/08) played by 7 different teams. For each class there are 25 clips of variable lengths, from a minimum of about 4 sec (corresponding to ∼ 100 frames) to a maximum of about 10 sec (∼ 2500 frames). This collection is particularly challenging because events are performed in a wide range of scenarios (i.e. different lighting conditions and different stadiums) and event classes show an high intra-class variability, because even instances of the same event may have very different progression. For our experiments videos are grouped in training and testing sets, selecting for each class 15 and 10 videos respectively, and results are obtained by 3-fold cross-validation.

The second dataset is composed by a subset of the TRECVID 2005 video corpus. It is obtained selecting five classes related to a few LSCOM dynamic concepts [21]. In particular we have selected the following classes: *Exiting Car*, *Running*, *Walking*, *Demonstration or Protest* and *Airplane Flying*. Examples of these events are shown in Fig. 5. The resulting video collection consists of about 180 videos for each class (∼ 860 in total) in MPEG-1 format with resolution 352 × 240 pixels and 30 fps. For each class, also in this dataset, videos have different lengths and show an high intra-class variability. Examples of high intra-class variability are shown in Fig. 6. Experiments are performed applying 3-fold cross-validation also for this dataset.

In the first experiment we have evaluated the effect on the classification accuracy of the metrics presented in Section 3 and of the codebook sizes. The second

[1] http://www.micc.unifi.it/vim

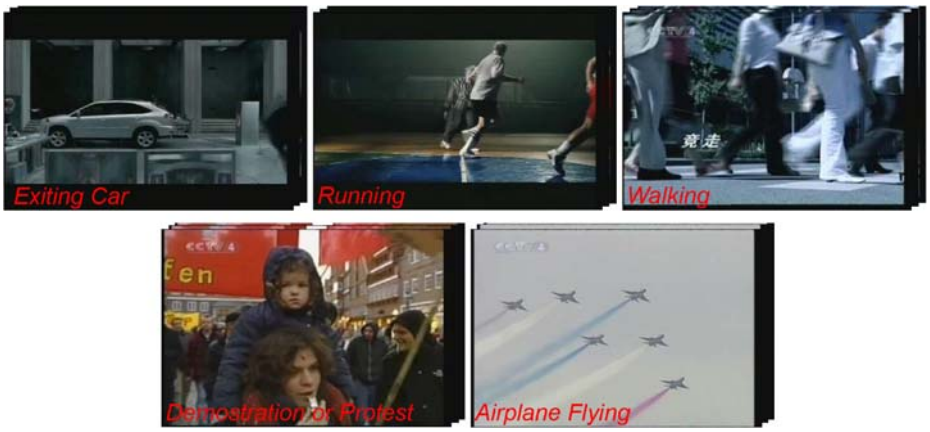


Fig. 5 Dataset based on a subset of the TRECVID 2005 video corpus. It consists of five events: Exiting Car, Running, Walking, Demonstration or Protest and Airplane Flying

experiment shows the improvement obtained using SVMs, based on the proposed string kernels, with respect to the baseline kNN classifier. In particular we used the LIBSVM implementation [9] using the "one-against-all" approach for multiclass classification. Finally, in the last experiment, we show that our method outperforms the traditional keyframe-based BoW approach.

(a) Running

(b) Exiting Car

Fig. 6 Examples of intra-class variability in two TRECVID 2005 classes: **a** shows two sequences containing the *Running* action, **b** shows two sequences containing an *Exiting Car*

5.1 Experiment 1: characters distance and codebook size

In this experiment we evaluate what is the best *characters* distance that has to be used when computing the Needleman-Wunsch distance and the best codebook size. It has been conducted on the soccer dataset, using a kNN classifier, varying the number of visual words used to build the codebook (from 30 to 500 codewords) and the metric used to compare the *characters* of the strings that represent video shots. Classification performances shows that the best codebook size is 200, while the best distance is the *Chi-square test*, since it has a more uniform performance, for the various classes of events, that is not achieved by the others (e.g. correlation metric). Table 1 reports the best results obtained for each distance, with a codebook of 200 words, along with the corresponding threshold. For these reasons we select *Chi-square* as the metric used in all the following experiments, and we set to 200 the codebook size for the soccer domain.

It can be observed in Table 2 that, unlike the case of object classification, the increase of the codebook size does not improve the performance and, instead, the effect may become negative. This can be explained by analyzing the type of views of the soccer domain: events are shown using the main camera that provides an overview of the playfield and of the ongoing event, and thus the SIFT points are mostly detected in correspondence of playfield lines, crowd and players' jerseys and shorts, as shown in Fig. 7, and thus the whole scene can be thoroughly represented using an histogram with a limited number of bins for the interest points. Increasing the number of bins may risk to amplify the intra-class variability and then reduce the accuracy of classification, resulting finally also in higher computational costs.

5.2 Experiment 2: comparison with kNN classifier

In this experiment we have compared the results of the baseline kNN classifier with the results of the SVM classifier using the proposed kernel on the soccer dataset. The mean accuracy obtained by the SVM (0.73) largely outperforms that obtained using the kNN classifier (0.54). Figure 8 reports the global accuracy and the confusion matrices for the kNN and SVM classifiers, respectively. A large part of the improvement, in terms of accuracy, is due to the fact that the SVM has a better performance on the two most critical actions: *shot-on-goal* and *throw-in*. This latter class has the worst classification results, due to the fact that it has an extremely large variability in the part of the action that follows immediately the throw of the ball (e.g. the player may choose several different directions and strengths for the throw, the defending team may steal the ball, etc.).

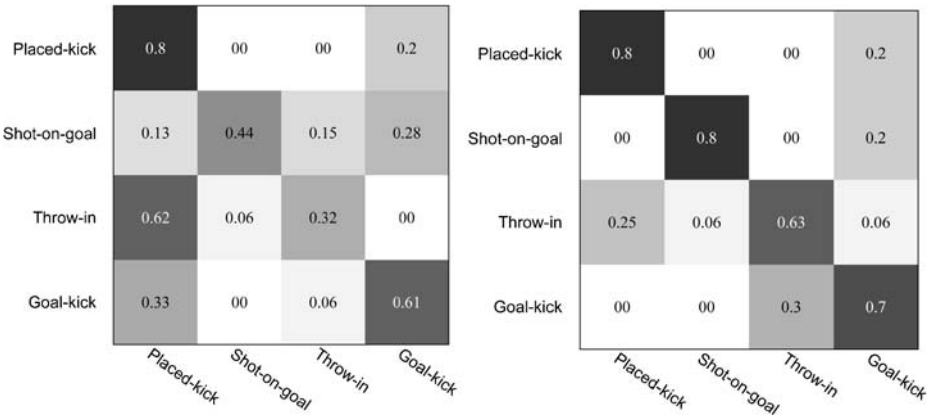
Table 1 Comparison of different metrics used to compare the *characters* (frequency vectors) of the strings that represent video shots

Metric	Th	Accuracy
Bhattacharyya	0.5	0.47
Chi-square	**0.13**	**0.54**
Correlation	0.7	0.53
Intersection	0.1	0.52
Kolmogorov-Smirnov	0.5	0.50
Mahalanobis	7	0.37

Table 2 Comparison of classification accuracy obtained using different codebook sizes on the soccer dataset

Codebook size	Accuracy
30	0.52
60	0.52
100	0.53
200	0.54
300	0.51
500	0.48

Fig. 7 Examples of SIFT points detected in a soccer video frame



	kNN	SVM
Mean Accuracy	0.54	**0.73**

(c)

Fig. 8 Confusion matrices of baseline kNN and the proposed SVM string classifiers; mean accuracy for kNN is equal to 0.54 and 0.73 for SVM with string kernel. **a** kNN classifier. **b** SVM string classifier. **c** Global accuracy

Table 3 Mean Average Precision (MAP) for event recognition in TRECVID 2005

	Exiting car	Running	Walking	Demonstration or protest	Airplane flying	**MAP**
BoW	0.25	**0.57**	0.28	0.32	0.17	0.32
Our approach	**0.37**	0.36	**0.29**	**0.38**	**0.34**	**0.35**

5.3 Experiment 3: comparison with a traditional keyframe-based BoW approach

Finally, in this experiment we show the improvement of the proposed method with respect to a traditional keyframe-based BoW approach, using the TRECVID dataset. As in the first experiment, we have initially tested different vocabulary sizes, looking for the correct choice for the TRECVID 2005 videos corpus. Results show that, in this case, a vocabulary of 300 words is a good trade-off between discriminativity and generalizability.

For a direct comparison we evaluate the classification performance using the Mean Average Precision (MAP) measure, which is the standard evaluation metric employed in the TRECVID benchmark. In particular, this measure gives a single numerical figure to represent the accuracy of a ranked concept detection result. Formally, let T be the size of the test set, R the number of relevant shots and R_i the number of relevant shot in the top i shot of a query result. $C_i = 1$ if the i^{th} shot is relevant and 0 otherwise. The average precision is defined as:

$$AP = \frac{1}{R} \sum_{i=1}^{S} \frac{R_i}{i} C_i. \tag{12}$$

MAP is the mean of average precision scores over a set of queries.

Table 3 reports the comparison results between a traditional BoW approach, as reported in [40], and the proposed method in terms of Mean Average Precision. Our method outperforms the traditional bag-of-words approach in four classes out of five, with an average improvement of 3%. We found a drop in classification performances only for the *Running* event. This is due to the fact this class shows a very high intra-class variability (see Fig. 6), with large differences in shot lengths. In particular, there are many different kinds of running actions in the dataset, each of which is depicted from a different camera viewpoint; for example, in several videos, often related to commercials, the running person is filmed frontally, while in many others people is filmed from the sides (e.g. in sports videos).

6 Conclusions

In this paper we have presented a method for event classification based on the BoW approach. The proposed system uses generic static visual features (SIFT points) that represent the visual appearance of the scene; the dynamic progression of the event is modelled as a *phrase* composed by the temporal sequence of the bag-of-words histograms (*characters*). Phrases are compared using the Needleman-Wunsch edit distance and SVMs with a string kernel have been used to deal with these feature vectors of variable length. Experiments have been performed on soccer videos and

TRECVID 2005 news videos; the results show that SVM with string kernels outperform both the performance of the baseline kNN classifiers and of the standard BoW approach and, more generally, they exhibit the validity of the proposed method. Our future work will deal with the application of this method to a broader set of events and actions that are part of the TRECVID LSCOM events/activities list, and the use of other string kernels. Moreover, we will investigate the possibility to integrate the proposed approach in an ontology-based framework [3], that exploits concept and event dependencies to improve the quality of classification.

Acknowledgements This work is partially supported by the EU IST VidiVideo Project (Contract FP6-045547) and IM3I Project (Contract FP7-222267). The authors thank Filippo Amendola for his support in the preparation of the experiments.

References

1. Bahlmann C, Haasdonk B, Burkhardt H (2002) On-line handwriting recognition with support vector machines—a kernel approach. In: Proc. of int'l workshop on frontiers in handwriting recognition
2. Ballan L, Bertini M, Del Bimbo A, Serra G (2009) Action categorization in soccer videos using string kernels. In: Proc. of IEEE int'l workshop on content-based multimedia indexing (CBMI). Chania, Crete
3. Ballan L, Bertini M, Del Bimbo A, Serra G (2009) Semantic annotation of soccer videos by visual instance clustering and spatial/temporal reasoning in ontologies. Multimedia Tools and Applications (in press)
4. Berg C, Christensen JPR, Ressel P (1984) Harmonic analysis on semigroups. Springer, Berlin
5. Bertini M, Del Bimbo A, Serra G (2008) Learning rules for semantic video event annotation. In: Proc. of int'l conference on visual information systems (VISUAL)
6. Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE Trans Pattern Anal Mach Intell 23(3):257–267
7. Boiman O, Irani M (2007) Detecting irregularities in images and in video. Int J Comput Vis 74(1):17–31
8. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. In: Proc. of ACM int'l workshop on computational learning theory
9. Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm
10. Chen J, Ye J (2008) Training svm with indefinite kernels. In: Proc. of int'l conference on machine learning (ICML)
11. Cover T (1965) Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. IEEE Trans Electron Comput 14(3):326–334
12. Dollar P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: Proc. of int'l workshop on VS-PETS
13. Ebadollahi S, Xie L, Chang SF, Smith JR (2006) Visual event detection using multi-dimensional concept dynamics. In: Proc. of IEEE int'l conference on multimedia and expo (ICME)
14. Fergus R, Perona P, Zisserman A (2003) Object class recognition by unsupervised scale-invariant learning. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
15. Fergus R, Perona P, Zisserman A (2005) A sparse object category model for efficient learning and exhaustive recognition. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
16. Francois ARJ, Nevatia R, Hobbs JR, Bolles RC (2005) VERL: an ontology framework for representing and annotating video events. IEEE Multimed 12(4):76–86
17. Gill PE, Murray W, Wright MH (1981) Practical optimization. Academic, London
18. Haasdonk B (2005) Feature space interpretation of svms with indefinite kernels. IEEE Trans Pattern Anal Mach Intell 27(4):482–492
19. Haubold A, Naphade M (2007) Classification of video events using 4-dimensional time-compressed motion features. In: Proc. of ACM int'l conference on image and video retrieval (CIVR)

20. Ke Y, Sukthankar R, Hebert M (2005) Efficient visual event detection using volumetric features. In: Proc. of int'l conference on computer vision (ICCV)
21. Kennedy L (2006) Revision of LSCOM event/activity annotations, DTO challenge workshop on large scale concept ontology for multimedia. Advent technical report #221-2006-7, Columbia University
22. Laptev I (2005) On space-time interest points. Int J Comput Vis 64(2–3):107–123
23. Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
24. Leslie C, Eskin E, Weston J, Noble WS (2003) Mismatch string kernels for SVM protein classification. In: Proc. of int'l conference on neural information processing systems (NIPS)
25. Lodhi H, Saunders C, Shawe-Taylor J, Cristianini N, Watkins C (2002) Text classification using string kernels. J Mach Learn Res 2:563–569
26. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
27. Luss R, D'Aspremont A (2008) Support vector machine classification with indefinite kernels. In: Proc. of int'l conference on neural information processing systems (NIPS)
28. Mikolajczyk K, Schmid C (2004) Scale and affine invariant interest point detectors. Int J Comput Vis 60(1):144–152
29. Moreno PJ, Ho PP, Vasconcelos N (2003) A kullback-leibler divergence based kernel for svm classification in multimedia applications. In: Proc. of int'l conference on neural information processing systems (NIPS)
30. Navarro G (2001) A guided tour to approximate string matching. ACM Comput Surv 33(1): 31–88
31. Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443–453
32. Neuhaus M, Bunke H (2006) Edit distance-based kernel functions for structural pattern classification. Pattern Recogn 39(10):1852–1863
33. Niebles JC, Wang H, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. Int J Comput Vis 79(3):299–318
34. Riedel DE, Venkatesh S, Liu W (2008) Recognising online spatial activities using a bioinformatics inspired sequence alignment approach. Pattern Recogn 41(11):3481–3492
35. Sadlier DA, O'Connor NE (2005) Event detection in field sports video using audio-visual features and a support vector machine. EEE Trans Circuits Syst Video Technol 15(10):1225–1233
36. Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local SVM approach. In: Proc. of int'l conference on pattern recognition (ICPR)
37. Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, New York
38. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: Proc. of int'l conference on computer vision (ICCV)
39. Smeaton AF Over P, Kraaij W (2006) Evaluation campaigns and TRECVid. In: Proc. of ACM int'l workshop on multimedia information retrieval (MIR)
40. Wang F, Jiang YG, Ngo CW (2008) Video event detection using motion relativity and visual relatedness. In: Proc. of ACM int'l conference on multimedia (MM)
41. Xiang T, Gong S (2008) Incremental and adaptive abnormal behaviour detectionq incremental and adaptive abnormal behaviour detection. Comput Vis Image Underst 111:59–73
42. Xu D, Chang SF (2008) Video event recognition using kernel methods with multilevel temporal alignment. IEEE Trans Pattern Anal Mach Intell 30(11):1985–1997
43. Yang J, Hauptmann AG (2006) Exploring temporal consistency for video analysis and retrieval. In: Proc. of ACM int'l workshop on multimedia information retrieval (MIR)
44. Yang J, Jiang YG, Hauptmann AG, Ngo CW (2007) Evaluating bag-of-visual-words representations in scene classification. In: Proc. of ACM int'l workshop on multimedia information retrieval (MIR)
45. Zhang D, Perez DG, Bengio S, McCowan I (2005) Semi-supervised adapted HMMs for unusual event detection. In: Proc. of int'l conference on computer vision and pattern recognition (CVPR)
46. Zhang J, Marszałek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: a comprehensive study. Int J Comput Vis 73(2):213–238
47. Zhou X, Zhuang X, Yan S, Chang SF, Hasegawa-Johnson M, Huang T (2008) Sift-bag kernel for video event analysis. In: Proc. of ACM int'l conference on multimedia (MM)

Lamberto Ballan received the MS degree in computer engineering in 2006 from the University of Florence, Italy, where he is currently a PhD student at the Visual Information and Media Lab at Media Integration and Communication Center. His main research interests focus on Multimedia Information Retrieval, Computer Vision and related fields such as Pattern Recognition and Machine Learning.

Marco Bertini is Assistant Professor at the Department of Systems and Informatics at the University of Florence, Italy. He received a MS in electronic engineering from the University of Florence in 1999, and PhD in 2004 from the same University. His main research interest is content-based indexing and retrieval of videos and Semantic Web technologies.

Alberto Del Bimbo is Full Professor of Computer Engineering at the University of Florence, Italy. He is also the Director of the Master in Multimedia, and the President of the Foundation for Research and Innovation at the same university. His scientific interests are Pattern Recognition, Image Databases, Human Computer Interaction and Multimedia applications. Prof. Del Bimbo is the author of over 230 publications in the most distinguished international journals and conference proceedings. He is the Associate Editor of Pattern Recognition, Journal of Visual Languages and Computing, Multimedia Tools and Applications, Pattern Analysis and Applications, and International Journal of Image and Video Processing, and was the Associate Editor of IEEE Transactions on Multimedia, and IEEE Transactions on Pattern Analysis and Machine Intelligence.

Giuseppe Serra received the laurea degree in computer engineering from the University of Florence in 2006. He is a PhD student at the Visual Information and Media Lab at the Media Integration and Communication Center, University of Florence. His research interests focus on Video Understanding based on Statistical Pattern Recognition and Ontologies, Multiple View Geometry, Self-calibration and 3D Reconstruction.