Video Event Classification Using Bag of Words and String Kernels

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, and Giuseppe Serra

Media Integration and Communication Center, University of Florence, Italy
{ballan,bertini,delbimbo,serra}@dsi.unifi.it

Abstract. The recognition of events in videos is a relevant and challenging task of automatic semantic video analysis. At present one of the most successful frameworks, used for object recognition tasks, is the bag-of-words (BoW) approach. However this approach does not model the temporal information of the video stream. In this paper we present a method to introduce temporal information within the BoW approach. Events are modeled as a sequence composed of histograms of visual features, computed from each frame using the traditional BoW model. The sequences are treated as strings where each histogram is considered as a character. Event classification of these sequences of variable size, depending on the length of the video clip, are performed using SVM classifiers with a string kernel that uses the Needlemann-Wunsch edit distance. Experimental results, performed on two datasets, soccer video and TRECVID 2005, demonstrate the validity of the proposed approach.

Keywords: video annotation, action classification, bag-of-words, string kernel, edit distance.

1 Introduction and Related Works

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 the problem of occlusions and geometrical transformations [1,2]. These approaches are commonly based on the idea of modeling a complex object or a scene by a collection of local salient points. Each of these local features describes a small region around the interest point and therefore they are robust against occlusion and clutter. In particular, SIFT features by Lowe [3] have become the de facto standard because of their high performances and relatively low computational cost. In fact, these features have been frequently and successfully applied to many different tasks such as object or scene recognition.

In this field, an approach that recently has become very popular is the Bag-of-Words (BoW) model. 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 word and it is represented by a bag of quantized invariant local descriptors

(usually SIFT), called *visual-words* or *visterms*. The main reason of 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 [4,5,6,7], like classification of objects or scenes, instead of recognizing a specific class of objects. The efficacy of this approach is demonstrated also by the large number of systems based on BoW representations that participate to the PASCAL VOC and TRECVID competitions. More recently, part-based and BoW models have been successfully applied also to the classification of human actions [8,9] and to video event recognition, typically using salient features that represent also temporal information (such as spatio-temporal gradients). These tasks are particularly interesting for video indexing and retrieval where dynamic concepts occur very frequently. Unfortunately, for this purpose the standard BoW model 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. Recently, few works have been proposed to cope with this problem. Wang *et al.* [10] have proposed to extend the BoW representation constructing relative motion histograms between visual words. In this way, they are able to describe motion of visual words obtaining better results on video event recognition. Xu *et al.* [11] represented each frame of video clips as a bag of orderless descriptors, applying then Earth Mover's Distance to integrate similarities among frames from two clips. They further build a multi-level temporal pyramid, observing that a clip is usually comprised of multiple sub-clips corresponding to event evolution over time; finally, video similarity is measured by fusing information at different levels.

In this paper, we present an approach to model actions as a sequence of histograms (one for each frame) represented by a traditional bag-of-words model. An action is described by a "phrase" of variable size, depending on the clip's length, providing so 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 [12] because it performs a global alignment on sequences dealing with video clips of different lenghts. Using this kind of representation we are able to perform categorization of video events and, following the promising results obtained in text categorization [13] and in bioinformatics (e.g. protein classification) [14], we investigate the use of SVMs based on an edit-distance based string kernel to perform classification. Experiments have been performed on soccer and news video datasets, comparing the proposed approach to a baseline kNN classifier and to a traditional BoW model. 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: the techniques for frame and action representation are discussed in Sect. 2; the classification method, including details about the SVM string kernel, is presented in Sect. 3; experimental results are discussed in Sect. 4 and, finally, conclusions are drawn in Sect. 5.

2 Action Representation

Structurally an action is a sequence of frames, and may have different lengths depending on how the action has been carried out. We represent an action by a sequence of visual words frequency vectors, computed from the frames of the sequence (Fig. 1); we call this sequence (string) *phrase*, where each frequency vector is considered as a *character*.
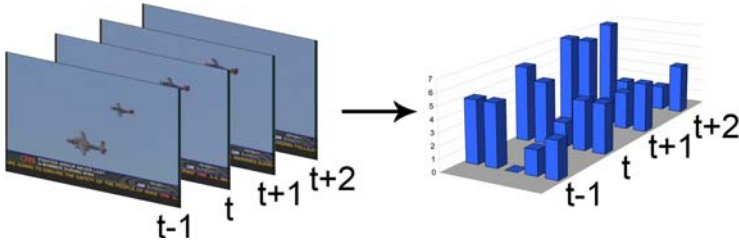

Fig. 1. Video clips are represented as a sequence of BoW histograms; actions are so described by a *phrase* (string) of variable size, depending on the clip's length

2.1 Frame Representation

Video frames are represented using bag-of-words, because this representation has demonstrated to be flexible and successful for various image analysis tasks [4,5,7]. First of all, a visual vocabulary is obtained by vector quantization of large sets of local feature descriptors. It is generated by clustering the detected keypoints (DoG in our case) in the feature space and using each cluster as a 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 model. A small vocabulary may lack the 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 detemined by experiments [6]. Once a vocabulary is defined, a visual word frequency vector is computed for each frame, counting the number of occurrences of each visual word of the vocabulary in that frame. This frequency vector is used as frame representation and it can be fed to a classifier for classification. In this work we use SIFT features [3] as local salient points and k-means clustering for vocabulary formation.

2.2 Action Representation

As previously introduced, each video clip 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.
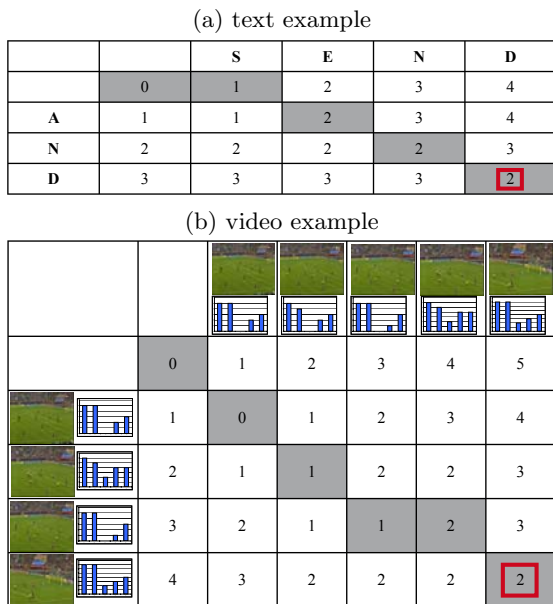
(a) text example

		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. 2. Needleman-Wunsch edit distance: (a) text and (b) video examples

Edit distance. The edit distance between two string of characters is the number of operations required to transform one of them into the other (substitution, insertion and deletion). In particular our approach uses the Needleman-Wunsch distance [12] because it 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 is to build up the best alignment through optimal alignments of smaller subsequences, using dynamic programming. 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:

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

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. Fig. 2 show 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 alignement between the sequences is highlighted in each cost matrix. The algorithm is $O(mn)$ in time and $O(min(m,n))$ in space, where m and n are the lengths of the two strings being compared.

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 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*.

3   Action Categorization

In the latest years Support Vector Machines (SVMs), introduced by Vapnik *et al.* [15], 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 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 classification that can be used to classify new input data. In the two classes classification problem suppose to have a training data set that comprises N input vectors $x_1, ..., x_N$, with corresponding target values $t_1, ..., t_N$ where $t_n \in \{-1, 1\}$. The SVMs approach finds the linear decision boundary $y(x)$ as:

$$y(x) = w^T \phi(x) + b \tag{1}$$

where ϕ denotes a fixed feature-space transformation, b is a bias parameter, so that, if the training data set is linearly separable, $y(x_n) > 0$ for points having $t_n = +1$ and $y(x_n) < 0$ 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:

$$\widetilde{L}(\mathbf{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m t_n t_m k(x_n, x_m) \tag{2}$$

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

$$\sum_{n=1}^{N} a_n t_n = 0, \quad a_n \geq 0 \quad \text{for} \quad n = 1, ..., N. \tag{3}$$

where $k(x_n, x_m)$, called kernel function, is defined by $k(x, x') = \phi(x)^T \phi(x')$. The parameters w and b are then derived from the optimal $\mathbf{a}$. The dual problem takes the form of a quadratic programming problem, which can be efficiently solved and any solution is a global optimum. Moreover, the SVM approach permits to use kernel techniques, so that the maximum margin classifier can be

found efficiently in a feature space whose dimensionality exceeds the number of data points. Recently, many approaches in image categorization have successfully used different kernels such as linear, radial and chi-square basis functions; in particular the latter gives the best results. However these kernels are not appropriate for action classification. In fact these kernels deal with input vectors with fixed dimensionality, whereas action representation vectors usually have different lengths depending on how it is performed. Unlike other approaches that solve this problem simply by representing the clips with a fixed number of samples [16], 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:

$$k(x, x') = exp(- \left\| x - x' \right\|^2 / 2\sigma^2). \tag{4}$$

we replace the Euclidean with the Needlmann-Wunsch distance. The proposed kernel is:

$$k(x, x') = exp(-d(x, x')). \tag{5}$$

where $d(x, x')$ is the Needlmann-Wunsch distance between x, x' input vectors. In this approach the structure of the string is evaluated by the edit distance and not by the kernel, that uses only the value of this distance. It has been demonstrated [17] that this type of kernels is suitable for classification of shapes, handwritten digits and chromosome images, despite the fact that the general edit distance has not been proved to be a valid kernel; this is confirmed in our experiments where all the pre-computed string kernels were checked to confirm their validity.

4   Experimental Results

We have carried out video event classification experiments to evaluate the general applicability and analyse the performance improvements achievable by the proposed method w.r.t. baseline kNN classifier and the standard BoW approach using a soccer videos and a subset of TRECVID 2005 video corpus. In the following sections, 4.1 and 4.2, the experiments and the two datasets used are described in details.

4.1   Comparing String-Kernel SVM Classifiers to Baseline kNN Classifiers on Soccer Videos Dataset

In this experiment we have compared the results of the proposed method with the baseline kNN classifier on a soccer video dataset. This dataset, available on request at our webpage[1], consists of 100 video clips in MPEG2 format at full PAL resolution (720 × 576 pixels, 25 fps). It contains 4 different actions: *shot-on-goal*, *placed-kick*, *throw-in* and *goal-kick*. The sequences were taken from 5 different matches of the Italian *"Serie A"* league (season 2007/08) between 7 different teams. For each class there are 25 clips of variable lengths, from a

---

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

minimum of ~ 4 sec (corresponding to ~ 100 frames) to a maximum of ~ 10 sec (~ 2500 frames). This collection is particularly challenging because actions are performed in a wide range of scenarios (i.e. different lighting conditions and different stadiums) and action classes show an high intra-class variability, because even instances of the same action may have very different progression. Videos are grouped in training and testing sets, composed by 20 and 5 videos respectively, and results are obtained by 3-fold cross-validation.

Results. Initially we have evaluated how different sizes of the visual vocabulary (30, 150, 300 visual words) affect the classification accuracy, obtaining the best result ($\sim 52\%$) with 30 words. In our test we observe that the increase of the codebook size does not improve the performance. This can be explained by analysing the type of views of the sport domain: actions are shown using the main camera that provides an overview of the playfield and of the ongoing action; thus the SIFT points are mostly detected in correspondence of playfield lines, crowd and players' jerseys and trousers and the whole scene can be completely represented using an histogram with a limited number of bins for the interest points. Increasing the number of bins risks to amplify the intra-class variability, even reducing the accuracy of classification, resulting also in higher computational costs. In another test we have evaluated what is the best metric to compare the characters (frequency vectors) and we have obtained the best accuracy using the Chi-square distance. Using the best dictionary size and metric selected with the previous tests we have finally compared the baseline kNN classifier and the proposed SVM with string kernel. The mean accuracy obtained by the SVM (0.73) largely outperforms that obtained using the kNN classifier (0.52). Fig. 3 reports the confusion matrices for 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.).

4.2 Comparing the Proposed Approach to a Baseline ("traditional") Bag-of-Words Representation on TRECVID 2005

In this experiment we show the improvement of the proposed approach with respect to a traditional BoW model. Experiments are performed on a subset of the TRECVID 2005 video corpus, obtained selecting five classes related to a few LSCOM dynamic concepts. In particular we have selected the following classes: *Exiting Car*, *Running*, *Walking*, *Demonstration Protest* and *Airplane Flying*. The resulting video collection consists of ~ 180 videos for each class (~ 860 in total); experiments are performed again applying 3-fold cross-validation.

Results. As in the previous experiment, we have initially experimented different vocabulary sizes looking for the correct choice in this video domain. Results show that, in this case, a vocabulary of 300 words is a good trade-off between
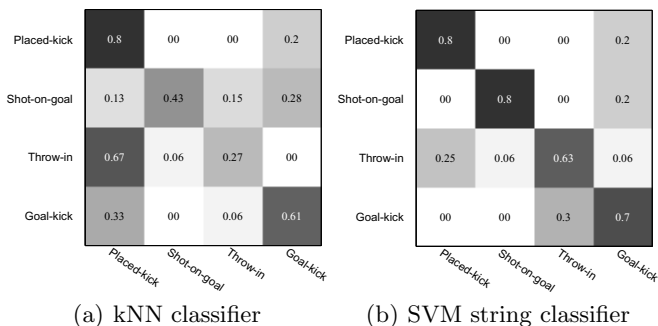
(a) kNN classifier   (b) SVM string classifier

Fig. 3. Confusion matrices of baseline kNN and SVM string classifiers; mean Accuracy for kNN is equal to 0.52 and 0.73 for SVM with string kernel

discriminativity and generalizability. Even in this case the metric used for comparing the similarity among characters within the N-W edit distance is the Chi-square (with a threshold of 4.5). Table 1 reports the comparison results between a traditional BoW aprroach and the proposed method; results are expressed in terms of Mean Average Precision (MAP) because it is the standard evaluation metric in the TRECVID benchmark.

Table 1. Mean Average Pracision (MAP) for event recognition in TRECVID 2005

	Exiting Car	*Running*	*Walking*	*Demo. Protest*	*Airplane Flying*	**MAP**
BoW	0.2	**0.49**	0.27	0.28	0.15	0.28
Our Approach	**0.37**	0.36	**0.29**	**0.38**	**0.34**	**0.35**

Our approach, on average slightly outperforms the traditional bag-of-words model (+7%) and it is also outperforming on four classes out of five.

5   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. 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 and 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 approach 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.

References

1. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.: A comparison of affine region detectors. International Journal of Computer Vision 65(1-2) (2005)
2. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(10) (2005)
3. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
4. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Proc. of ICCV (2003)
5. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: Proc. of CVPR (2003)
6. Yang, J., Jiang, Y.G., Hauptmann, A.G., Ngo, C.W.: Evaluating bag-of-visual-words representations in scene classification. In: Proc. of ACM MIR (2007)
7. Zhang, J., Marszałek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: A comprehensive study. International Journal of Computer Vision 73(2), 213–238 (2007)
8. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: Proc. of VSPETS (2005)
9. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised Learning of Human Action Categories Using Spatial-Temporal Words. International Journal of Computer Vision 79(3), 299–318 (2008)
10. Wang, F., Jiang, Y.-G., Ngo, C.-W.: Video event detection using motion relativity and visual relatedness. In: Proc. of ACM Multimedia (2008)
11. Xu, D., Chang, S.-F.: Video event recognition using kernel methods with multi-level temporal alignment. IEEE Transactions on Pattern Analysis and Machine Intelligence 30(11) (2008)
12. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology 48(3), 443–453 (1970)
13. Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., Watkins, C.: Text classification using string kernels. Journal of Machine Learning Research (2002)
14. Leslie, C., Eskin, E., Weston, J., Noble, W.S.: Mismatch string kernels for SVM protein classification. In: Proc. of NIPS (2003)
15. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proc. of ACM Workshop on Computational Learning Theory (1992)
16. Sadlier, D.A., O'Connor, N.E.: Event detection in field sports video using audio-visual features and a support vector machine. IEEE Transactions on Circuits and Systems for Video Technology 15(10), 1225–1233 (2005)
17. Neuhaus, M., Bunke, H.: Edit distance-based kernel functions for structural pattern classification. Pattern Recognition 39(10), 1852–1863 (2006)