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 (MICC), University of Florence, Italy
http://www.micc.unifi.it/ballan - ballan@dsi.unifi.it

1. Introduction

***Goal:* define an effective representation to perform video event classification.**

- 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.
- An approach that has become very popular is the **Bag-of-Words (BoW) model** - originally proposed 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 is represented by **a bag of quantized invariant local descriptors** (usually SIFT), called *visual-words*.
- More recently, it has been successfully applied also to the **video event classification problem**, and the most common solution is to apply the traditional BoW approach using static features (e.g. SIFT) on a keyframe basis.
- To this end, the standard BoW model has shown some drawbacks; the most evident problem is that it **does not** take into account **temporal relations** between consecutive frames (see Fig. 1).

(a) (b) (c)

Figure 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?*

2. Approach

- Actions and events are modeled 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 length, depending on the clip's duration, thus providing a global description of the video content that is able to **incorporate temporal relations**.

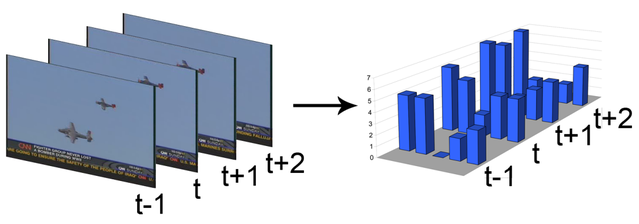


Figure 2: *A video shot represented as a sequence of BoW histograms.*

- Video phrases can be compared by computing **edit distances** between them; we apply the **Needleman-Wunsch** (NW) distance because it performs a global alignment on sequences dealing with video clips of different lengths.
- Finally, we investigate the use of **SVMs with a string kernel**, based on NW edit distance, to perform classification.

3. Framework

An overview of our approach is illustrated in Fig. 3.

- In the training stage the features (SIFT) extracted from videos are clustered using k-means into visual words (A,B,C,D,E); each video is represented as a sequence of BoW histograms.
- 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.

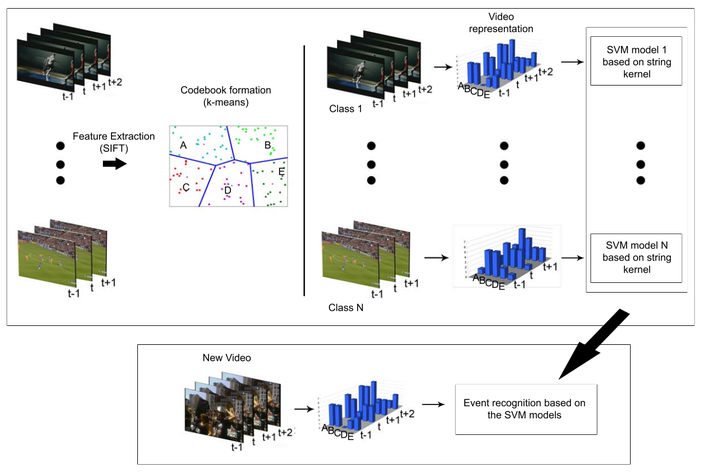


Figure 3: *Schematization of the approach.*

4. Edit distance

Definition *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).*

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

(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

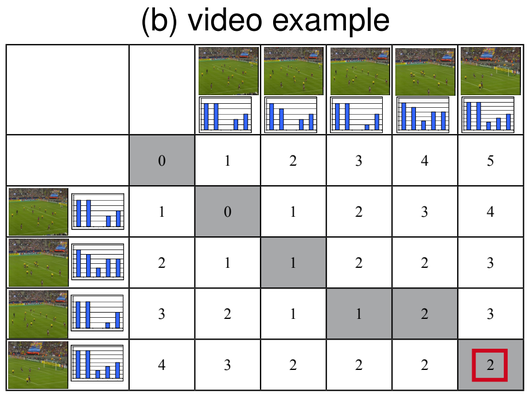


Figure 4: *Needleman-Wunsch edit distance: (a) text and (b) video examples.*

Measuring similarity between characters. A crucial point is the evaluation of the similarity among characters ($a_i \approx b_j$).

- In the text case, the number of characters is limited and it permits to define a similarity matrix between characters.
- In our case each frequency vectors is a different character; this requires to define a function that evaluates the similarity of two characters to reduce the alphabet size (we apply *Chi-square*).

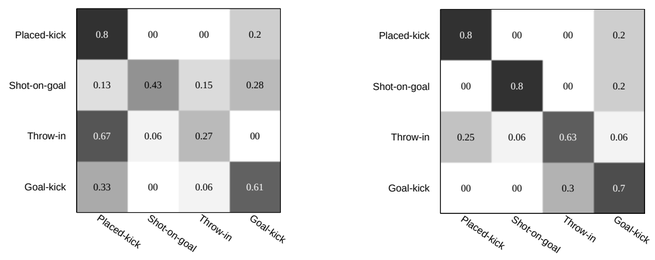
5. Results

Experiments have been performed on two different domains:

- **Soccer videos:** the dataset consists of 100 video clips in MPEG-2 format at full PAL resolution, and it contains 4 different events: *shot-on-goal*, *placed-kick*, *throw-in* and *goal-kick*.
- A subset of the **TRECVID 2005** news video corpus: it consists of 5 classes related to a few LSCOM dynamic concepts, *Exiting Car*, *Running*, *Walking*, *Demonstration or Protest* and *Airplane Flying*.

Figure 5: *TRECVID 2005 subset.*

Comparison with kNN classifier. We have compared the results of the baseline kNN classifier with the results of the SVM classifier using the proposed string kernel on the soccer dataset.



(a) kNN classifier (b) SVM string classifier

	kNN	SVM
Mean Accuracy	0.52	**0.73**

(c) Global accuracy

Figure 6: *Classification results of kNN and SVM string classifiers on soccer dataset.*

Comparison with the traditional BoW model. In this experiment we show the improvement of the proposed approach with respect to the standard keyframe-based BoW model (Wang *et al.*, MM08), using the TRECVID dataset.

	Our Approach	Wang *et al.*, MM08
MAP	**0.35**	0.28

Our approach outperforms the traditional BoW model in 4 classes out of 5, with an average improvement of 7% in terms of Mean Average Precision (MAP).

References

[1] L. Ballan, M. Bertini, A. Del Bimbo, and G. Serra. Action categorization in soccer videos using string kernels. In *Proc. of IEEE Int'l Workshop on Content-Based Multimedia Indexing (CBMI)*, Chania, Crete, 2009.