Modeling Video Activity with Dynamic Phrases and its Application to Action Recognition in Tennis Videos

Jonathan Vainstein[1,2] José F. Manera[1,2] Pablo Negri[3]
Claudio Delrieux[1] Ana Maguitman[2]

[1] Laboratorio de Ciencias de las Imágenes (IIIE - CONICET)
Departamento de Ingeniería Eléctrica y de Computadoras (DIEC)
[2] Grupo de Investigación en Administración de Conocimiento y Recuperación de Información
Departamento de Ciencias e Ingeniería de la Computación (DCIC)
Universidad Nacional del Sur (UNS)
Av. Alem 1253, (B8000CBP), Bahá Blanca, Argentina
[3] Instituto de Tecnología
UADE-CONICET
Lima 717, Buenos Aires, Argentina

Abstract. We present a novel approach to action recognition in tennis shot sequences. The underlying model considers the per-frame motion to be regarded as a *word* (within an alphabet of possible motions), and the sequence of frames as a *phrase* whose meaning is determined by the words given in a specific order. This feature extraction mechanism allows a semantic treatment of the classification stage using Conditional Random Fields. The system was applied on the RGB videos of the THETIS dataset, achieving an accuracy of over 86% in recognizing 12 different tennis shots among several takes produced by 55 different amateur and professional players.

Keywords: Action recognition, conditional random fields, support vector machines, optical flow, motion description

1 Action Recognition in Tennis Videos

The widespread availability of digital videos enables several development fields that requires automatic or semi-automatic video action labeling in different domains, such as action detection in surveillance systems [15], traffic accidents [8], and sports videos [19]. Some of these problems require efficient video processing and action detection methods that could be implemented in real time. In video processing of sport applications, it could be highly desirable to embed more sophisticated algorithms in low-cost cameras, for instance to be able to catch relevant actions instantaneously.

In particular, action recognition in tennis videos is subject of extensive research, since it poses significant challenges like distinguishing among different kinds of shots (which is a difficult task even for humans). Two problems are faced in this particular domain: First, an adequate feature extraction is required to obtain a lean but distinctive representation of the player's movements. Second, a sensitive and specific classifier is needed to recognise a shot among a set of possibilities.

In [12] a system for automatic annotation of actions in tennis matches was developed. In this system, the positions of the player and the ball were used as features, and the player's movements were analyzed based on silhouette transitions. Hidden Markov models and 2D appearance based models were used to identify the specific action. A different approach was taken in [5], where the authors used a motion descriptor based on the optical flow of a space-time volume with a nearest neighbor classifier to perform actions classification. In a similar way, [18], combined optical flow feature at low-level with Support Vector Machines (SVMs) at high-level. In [11] the detection of the player's movements was carried out using the Mean-shift algorithm, and optical flow was used to model the player's movement patterns over the field, while Conditional Random Fields (CRFs) were used for action recognition.

In this work we present a novel action recognition method aimed to identfy different shots in tennis videos. Motion in the region of interest (ROI) between consecutive frames is split into different parts (typically a 3×3 grid), and the motion direction and speed in every part is quantized in a set of discrete values. Thus, a particular frame can be regarded as a *word* that represents the motion speed and direction at every part of the ROI. Thus, an action (a sequence of frames) can be represented as a *phrase*, whose meaning depends on the specific words and their relative order. This feature extraction mechanism has several advantages: First, it allows a very parsimonious action representation. Second, it is simple and economic to compute (a *word* of 18 letters suffices to represent a whole frame), and several extensions and variations are readily possible. Third, it allows a semantic treatment of the classification step, using classification methodologies better suited for semantic analysis, such as CRFs. CRFs can represent state-to-state and feature-to-state dependencies in a natural way. This allows to take advantage of the meaningful information present in the order itself (which frame occured before and after a specific frame), thus obtaining higher accuracy than other classifiers when applied to sequential classification problems.

The recognition system was applied on the RGB videos of the THETIS dataset [6]. Actions were represented as dynamic phrases, and those phrases were classified using CRFs, achieving an accuracy of over 86% in recognizing between 12 different tennis shots from 8374 takes performed by 55 different players, thus obtaining a performance that appears to be beyond the state-of-the-art proposals to cope with this action recognition problem. In addition, classification was also tested with Support Vector Machines (SVMs) to verify if there is indeed a significant accuracy gain in using CFRs. The rest of this paper is organised as follows. In the next two sections we describe respectively the feature extraction and classification mechanisms that were implemented. In Sect. 4 we present the main results of this contribution, and in Sect. 5 we elaborate on the conclusions and discuss further work.

2 Feature Extraction and Motion Representation

The feature extraction framework is composed of three stages: low level video processing, feature description, and dynamic phrase enconding. For the low level processing, we computed the pixelwise AND between the synchronized RGB videos, and the binary videos of the player's silhouettes (albeit other background removal algorithms can be

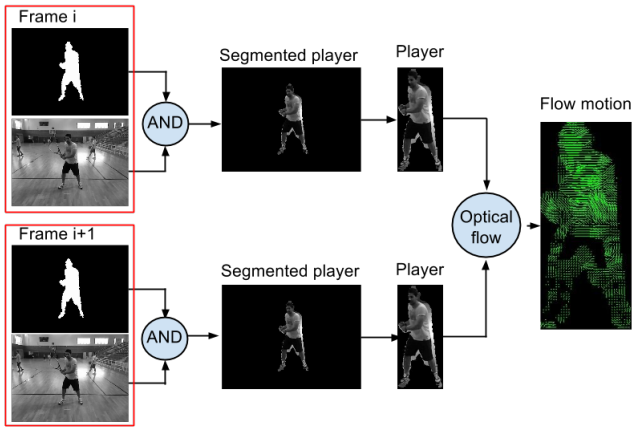

Fig. 1: Video Processing

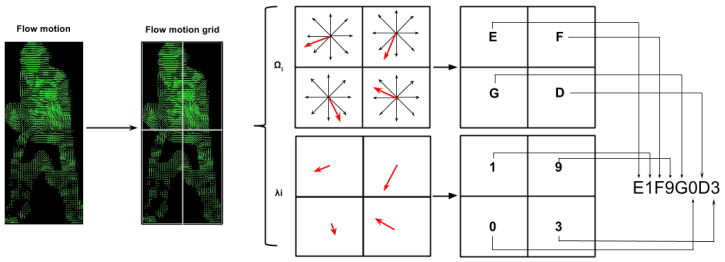

Fig. 2: Feature Extraction: the *dynamic word* 'E1F9G0D3' encodes the motion information of the frame.

used instead). The ROI of each frame is defined as the minimum bounding box containing the silhouette of the player, within which the optical flow between successive frames is computed (see Fig. 1).

Then, the optical flow in the ROI is divided into $q = M \times N$ equal-sized regions or cells. For each cell i of a given frame f, two features are computed:

- Ω_i, the predominant direction of the optical flow between 0 and 2π is equaly quantized into 8 discrete values, each represented with one of the first 8 uppercase letters (*i.e.*, A represents predominant flow between 0 and $\frac{\pi}{4}$ and so on).
- λ_i, the sum of the moduli of the displacement vectors of the optical flow in the cell is equaly quantized into 10 values (wrt the maximum observed).

Then, the *dynamic word* representing the motion on frame f is very concisely represented as a $2q$ feature vector d_f:

$$d_f = \Omega_0\lambda_0\Omega_1\lambda_1 \cdots \Omega_q\lambda_q, \tag{1}$$

where $\Omega_i \in \{A, B, C, D, F, E, G, H\}$ and $\lambda_i \in \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9\}$.

Figure 2 shows an example of the mapping described above. Finally, a Dynamic Phrase is the sequence of dynamic words comprising a given movement. Figure 3 shows an example of this encoding for a specific video sequence.

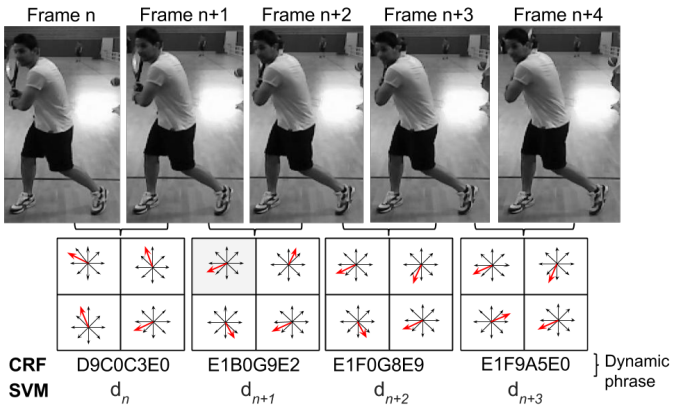

Fig. 3: Dynamic Phrase enconding of a sequence.

3 Classification Methodology

We tested the accuracy of our feature extraction schema with two different classifiers, CRFs and SVMs.

Conditional Random Fields

Our classification task can be regarded as a multivariate prediction problem, where we wish to predict a sequence $\mathbf{y} = \{y_0, y_1, \cdots, y_T\}$ of random variables given a sequence $\mathbf{x} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_T\}$ of feature vectors. A possible approach is to learn an independent per-position classifier that maps $\mathbf{x} \mapsto y_i$, for each $i \leq T$.

However, this approach does not capture the complex dependencies that may be present in the output variables. In order to overcome this limitation, the dependencies among output variables can be naturally represented by means of graphical models. These models represent a complex distribution over many variables as a product of local factors on smaller subsets of variables, based on independence assumptions suggested by knowledge of the domain.

An effective approach to learn in these models is to represent the conditional distribution $p(\mathbf{y}|\mathbf{x})$ based on these factors.

This class of statistical modeling method is known as Conditional Random Fields (CRFs) [9]. Although early applications of CRFs used linear chains, more general graphical models for predicting complex structures (e.g., graphs and trees) have also been proposed. In this work we focus on the use of linear-chain CRFs to model a simple form of dependency, in which the output variables are arranged in a sequence. This is useful for the task of action recognition in video, where input features are obtained from a sequence of frames.

A linear-chain CRF can naturally model the sequential dependencies between frames by defining the conditional probability for the sequence $\mathbf{y} = \{y_0, y_1, \cdots, y_T\}$ of output

variables given the sequence $\mathbf{x} = \{\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_T\}$ of observable random vectors as

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp\Big[\sum_{t=1}^{T} \phi_t(y_t, y_{t-1}, \mathbf{x}_t) + \phi_0(y_0, \mathbf{x}_0)\Big]$$

where $\phi_t(\cdot)$ is the local potential (or score) function of the model at time (or frame) t, and $Z(\mathbf{x})$ is a partition function that ensures that the conditional probability $p(\mathbf{y}|\mathbf{x})$ of a sequence $\mathbf{y}$ will sum to one. In our specific problem, each predicted variable y_t is a tennis shot in $L = \{$backhand, backhand2hands, backhand_slice, $\cdots\}$, with $|L| = 12$. Each observed feature vector $\mathbf{x_t}$ is instantiated with a dynamic word describing a specific frame, as illustrated in Figure 3. In this way, a CRF classifier can be naturally trained by means of dynamic phrases, where the goal is to learn the distribution $p(\mathbf{y}|\mathbf{x})$ over the tennis shot labels ($\mathbf{y}$) given the phrases describing each movement ($\mathbf{x}$).

Various methods can be used to train CRFs, including the penalized log-likelihood criteria, pseudo log-likelihood, voted perceptron, margin maximization, gradient tree boosting and logarithmic pooling (a description of these methods can be found in [7]). In this work we used CRFSuite [14] as a software tool to build and test the CRF models. Training was performed using a combination of the Limited-memory BFGS method [13] and the Orthant-Wise Limited-memory Quasi-Newton method [1].

Support Vector Machines

Actions recognition represents a particular challenge because sequences of the same movement have, in general, different number of frames. As a result, the complete set of descriptors computed on the sequences cannot be used for SVM classifiers, considering that these classifiers have a fixed number of inputs. In this work, we employ a methodology inspired on local features proposed by Wallraven et al. [16], where a non-linear SVM classifier is implemented using a kernel function $K_S(x, y)$ satisfying Mercer's theorem. We define $\mathcal{S} = \{\mathbf{S}_i\}_{i=1}^{m}$ as a set of shot sequences, $\mathcal{F} = \{\mathbf{F}_i\}_{i=1}^{m}$ as the corresponding set of dynamic phrases, with $\mathbf{F}_i = \{\mathsf{w}_j(\mathbf{S}_i)\}_{j=1}^{n_i}$, $i = 1, \ldots, m$, and $\mathsf{w}_j(\mathbf{S}_i)$ the j-th dynamic word of sequence $\mathbf{S}_i$. For SVM, dynamic words consist on the Cartesian representation if we consider the components of the dynamic phrase as vectors (see eq. 1). For the pair $\langle \mathbf{S}_h, \mathbf{S}_k \rangle$ in $\mathcal{S}$, the methodology seeks for the nearest words of $\mathbf{S}_h$ in $\mathbf{S}_k$, disregarding the position in the sequence. Alternatively, the kernel value is completed by searching on the other direction: looking from the phrase $\mathbf{S}_k$ to the phrase $\mathbf{S}_h$. All minimal distances are averaged in order to obtain a single value, measuring the similarity between both phrases. An extension on the local features model is applied on [16], by adding a coefficient which takes into account the position of the words in the phrase. The words are compared with [3] : $D_{a,b}(\mathsf{w}_{jh}, \mathsf{w}_{jk}) = \sum_i |\mathsf{w}^a_{jh,i} - \mathsf{w}^a_{jk,i}|^b$, using $a = 1$ and $b = 2$. Empirical results show better performance with this similarity function than with the popular χ^2 [17].

To evaluate a test sequence, it is necessary to compute its similarity with each training sample. This precomputed Kernel represents the input of the SVM classifier. Multiclass classification is solved using a *one-against-all* approach, selecting the class with the highest score. The SVM classification tests with the precomputed Kernel K_S were performed using LibSVM [2].

4 Experimental Setup and Results

THETIS contains 8374 video clips that consist on 12 different tennis shots, performed several times by 24 experienced and 31 unexperienced players. The shots were captured using Kinect, resulting in 5 different synchronized clips: RGB, silhouette, depth, 2D skeleton and 3D skeleton. This dataset provides an extra challenge to automated action recognition because while some shots may be distinguishable to an expert's eye, others can be quite confusing to someone who does not know in detail the different tennis shots. In [6], two methods were tested with this dataset: Space-Time Interest Points [10] and Dense Trajectories [17]. Both algorithms were applied on the 3D video clips, using skeleton and depth information.

In our experimental setup, the same strategy used in [6] was choosen for comparison purposes. This approach, called leave-one-person-out cross validation, preserves all the videos belonging to one subject as test set, while the others are used as training samples. This procedure is repeated N times, where N is the number of subjects within the dataset.

Two encodings were evaluated, one with a 2×2 grid (8 attributes per feature vector), and the other with a 3×3 grid (18 attributes). Figure 4 presents this classification results (using CRF and SVM), together with those reported in [6]. Using a 2×2 grid requires only 8 bytes to encode each frame, with an acceptable recognition accuracy. On the other hand, using a 3×3 grid requires 18 bytes per frame but the accuracy is significantly better. The confusion matrix obtained with CRFs and the 3×3 grid is shown in Figure 5, where the classifications among pairs of classes (predicted in columns, and actual in rows) are depicted in grayscale levels (darker meaning more frequent). The hit rate, *i.e.*, the *hits* of the classifier for every actual class, is shown in the corresponding diagonal element.

The recognition pipeline performs well within real-time constraints, *i.e.*, is able to recognise actions well within the framerate with standard desktop PCs and laptops. As a side-note, the recognition pipeline was implemented and tested on a Raspberry-Pi embedded system, and performance was still within the framerate.

Fig. 4: Classification results on the THETIS dataset: Space-time interest point (STIP) and dense trajectory data are taken from [6].

Methodology	Avg. Accuracy (%)
STIP - THETIS Depth	60.23
STIP - THETIS 3D Skeleton	54.40
Dense Trajectory - THETIS Depth	57.50
Dense Trajectory - THETIS 3D Skeleton	53.08
2x2 grid	
Dynamic Phrase - CRF - THETIS RGB	61.17
Dynamic Phrase - SVM - THETIS RGB	44.91
3x3 grid	
Dynamic Phrase - CRF - THETIS RGB	**86.44**
Dynamic Phrase - SVM - THETIS RGB	51.20

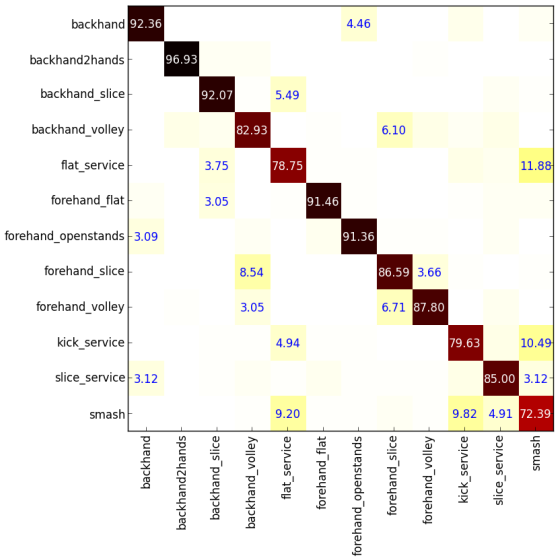

Fig. 5: CRF confusion matrix for a 3x3 grid, predicted class in columns, and actual class in rows.

It is arguable if the quantization in our feature representation produces a significant loss of information, and if this is so, if SVM with a full representation of the motion flow would achieve a better classification performance. The preliminary results presented here seem to contradict this view, which are also in agreement with the poor results presented in [6] with full fledged feature representation. The moral appears to be that CRFs are capturing essential information (*i.e.*, the actual order in the sequence) in an unique way, and that this information is more significant for recognition purposes than the exact motion values.

5 Conclusion, and Further Research

We proposed a novel action recognition methodology that combines the advantages of a lean action representation together with the discriminative power of CRFs. Our feature extraction mechanism represents actions as a sequence of words, each encoding the main movement speed and direction in different parts of the ROI. This representation is both parsimonious and economic to compute, and allows a more semantic treatment of the classification step. Finally, our methodology uses CRFs to take advantage of the sequential nature of this classification problem. The results appear to be more accurate than other proposals in similar contexts.

We are currently researching different optimizations on the ideas presented so far. Among them, it is sensible to use sliding windows [4] to represent a richer motion flow among n frames, taken every other m frames (given that successive frames will have a

high correlation). However, it is still a matter of experimental research to find optimal values for m and n. Also, it is unclear if the use of RGBD videos provides advantages in accuracy or performance. Depth information may improve tracking and background substraction, but 3D feature extraction for action modeling and representation in CRFs' may add noisy parameters that would actually deteriorate the classification accuracy.

References

1. G. Andrew and J. Gao. Scalable training of l1-regularized log-linear models. proceedings of the 24th international conference on machine learning. In *ICML*, pages 33–40, 2007.
2. C.-C. Chang and C.-J. Lin. Libsvm: A library for support vector machines. *ACM TIST*, 2(3):1–27, 2011.
3. O. Chapelle, P. Haffner, and V. Vapnik. Support vector machines for histogram-based image classification. *Neural Networks*, 10(5):1055–1064, 1999.
4. T. Dietterich. Machine learning for sequential data: A review. In *Structural, Syntactic, and Statistical Pattern Recognition*, volume 2396 of *LNCS*, pages 15–30. 2002.
5. A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, volume 2, pages 726–733, Washington, USA, 2003.
6. S. Gourgari, G. Goudelis, K. Karpouzis, and S. Kollias. Thetis: Three dimensional tennis shots a human action dataset. In *CVPRW*, pages 676–681, 2013.
7. R. Gupta. Conditional random fields. Unpublished report, ITT Bombay, 2006.
8. S. Kamijo, Y. Matsushita, K. Ikeuchi, and M. Sakauchi. Incident detection at intersections utilizing hidden markov model. In *ITS*, 1999.
9. J.D. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pages 282–289, 2001.
10. I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003.
11. F. Manera, J. Vainstein, C. Delrieux, and A. Maguitman. Action recognition in tennis videos using optical flow and conditional random fields. *AST JAIIO*, pages 152–162, 2013.
12. H. Miyamori and S. Iisaku. Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In *AFGR*, pages 320–325, 2000.
13. J. Nocedal. Updating quasi-newton matrices with limited storage. *Mathematics of Computation*, 35(151):773–782, 1980.
14. Naoaki Okazaki. Crfsuite: a fast implementation of conditional random fields (crfs), 2007.
15. M. Takahashi, M. Naemura, M. Fujii, and S. Satoh. Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories. In *CVPRW*, pages 9–16, 2011.
16. C. Wallraven, B. Caputo, and A. Graf. Recognition with local features: the kernel recipe. In *ICCV*, pages 257–264, 2003.
17. H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.
18. G. Zhu, C. Xu, W. Gao, and Q. Huang. Action recognition in broadcast tennis video using optical flow and support vector machine. In *ECCV*, pages 89–98, 2006.
19. G. Zhu, C. Xu, and Q. Huang. Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In *ACM*, pages 431–440, 2006.