Action Recognition in Tennis Videos using Optical Flow and Conditional Random Fields

José F. Manera[1,2] Jonathan Vainstein[1,2] 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 - LIDIA
Departamento de Ciencias e Ingeniería de la Computación (DCIC)

Universidad Nacional del Sur (UNS)
Av. Alem 1253, (B8000CBP), Baha Blanca, Argentina
Tel: (0291) 459-5135 / Fax: (0291) 459-5136

Abstract. The aim of Action Recognition is the automated analysis and interpretation of events in video sequences. As result of the applications that can be developed, and the widespread availability and popularization of digital video (security cameras, monitoring, social networks, among many other), this area is currently the focus of a strong and wide research interest in various domains such as video security, human-computer interaction, patient monitoring and video retrieval, among others. Our long-term goal is to develop automatic action identification in video sequences using Conditional Random Fields (CRFs). In this work we focus, as a case of study, in the identification of a limited set of tennis shots during tennis matches. Three challenges have been addressed: player tracking, player movements representation and action recognition. Video processing techniques are used to generate textual tags in specific frames, and then the CRFs are used as a classifier to recognise the actions performed in those frames. The preliminary results appear to be quite promising.

Key words: action recognition, tracking, conditional random fields, optical flow.

1 Introduction

The widespread availability and popularization of digital video makes necessary the development of automatic or semi-automatic systems for video action labeling in different domains, such as action detection in surveillance systems [1], traffic accidents [2] and sports videos [3].

This paper is focused on the analysis of video footage of tennis matches. One of the novel aspects of our proposal is the use of Conditional Random Fields (CRFs) instead of other classifiers. A CRF [4] [5] is an discriminative probabilistic model used to calculate the probability of a possible label sequence

conditioned on the observation sequence. This restricts the probabilistic model to the sequence of observations, thus avoiding the computation of probabilities for each possible sequence. Instead of relying on the joint probabilities $P(X, Y)$, the CRFs specify the probability of any given label sequences observation $P(Y|X)$. A typical graph of a linear-chain CRF model is illustrated in Fig. 1, where X and Y refer to observations and tag sequences respectively.

Graphical modeling is a powerful framework for representation and inference in multivariate probability distributions. Despite the fact that distributions over many variables can be expensive to represent, many of these distributions could be represented through graphical modeling as a product of local functions, where each function depends on a much smaller subset of variables. This factorization is related with certain conditional independence relationships among the variables.

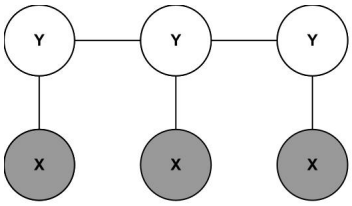


Fig. 1: Linear-chain CRF. The nodes tagged with X belong to observations and the nodes tagged with Y belong to labels.

Our long-term goal is the design and implementation of a system for action recognition in tennis videos. More specifically, we expect to combine the study of video processing techniques and CRFs with the aim of developing new techniques and achieving improvements in both areas.

There are several challenging tasks that need to be addressed in this line of research, including tracking, feature extraction, classification and pattern recognition. In addition, two challenges are faced for this particular application domain, namely the representation of the player's movements and action recognition. The detection of the player's movements is carried out using Mean-shift and optical flow [6] to model the movement patterns of the player in the field, while CRFs are used for action recognition. In this work, the CRFs take the information obtained from the optical flow of successive frames as the classifier input. The system is trained and tested with tennis video clips, which were manually labeled based on different kinds of tennis shots occurring in each one.

2 Related Work

Analysis of tennis videos is currently subject of widespread attention. In [7] a system for automatic annotation of actions in tennis matches is developed. In this system, player's and ball position are used as features and players' behaviors are analyzed based on silhouette transitions. Hidden markov model and 2-d

appearance based model are used to identify behavior category. A different approach is taken in [8], where the authors use a motion descriptor based on the optical flow of a space-time volume with a nearest neighbor classifier to perform actions' classification. In [9], optical flow is used as a low-level feature descriptor of the players movements, while Support Vector Machines were used to train the classifier using the optical flow information as input.

CRFs have been applied to a variety of domains, such as natural language processing [10] [11] [12] [13], bioinformatics [14] [15] and computer vision. In the latter area, some authors have used CRFs for labeling images [16]. In [17] CRFs are used to determine characteristics of parts of an object. In particular, for the case of recognition of actions in video sequences several works on detection of actions in sports videos [9] [18] were developed.

3 The Proposed Framework

As in any supervised machine learning process it is possible to distinguish two major stages. The first stage involves the construction of a model from labeled training data, which consists of a set of training examples. In the second stage the model is used to classify new examples. These two stages can be modeled by the two pipelines presented in Fig. 2 and described next.

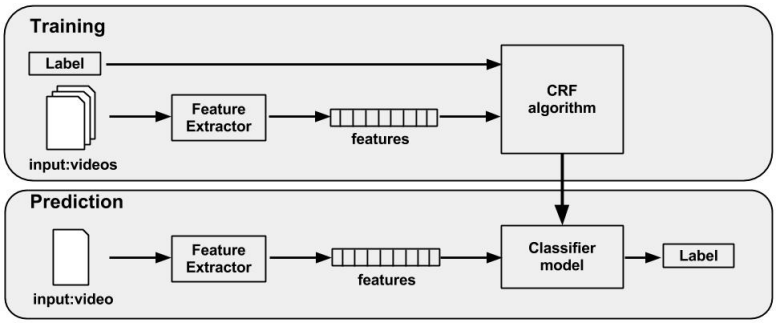


Fig. 2: Pipelines corresponding to training and classification.

3.1 Training

The training phase consists of a pipeline whose stages are: tracking, feature extraction and classifier parameter extraction. The videos used as input for this stage are of public domain and were taken with an oblique camera (Fig. 3). These videos were previously classified in two classes: right-swing and left-swing.

The first stage aims to perform the tracking of the player who is filmed from behind and selected by the user in the frame 0 of the video (Fig. 4). After selecting the region of interest (ROI) a player model is generated, which is represented

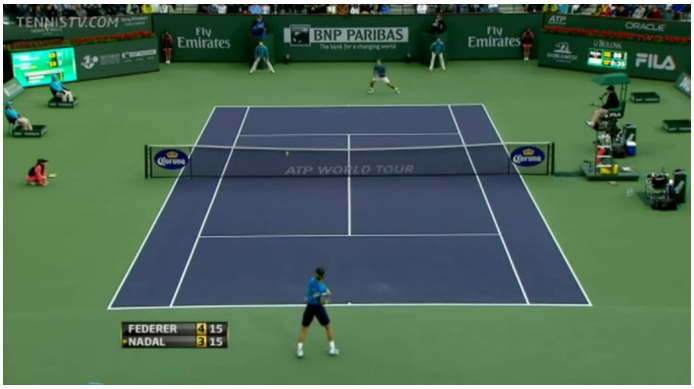


Fig. 3: Frame sample of a tennis video.

by two histograms. The first histogram consists of the values of luminance of the pixels corresponding to the player's clothes. This histogram is calculated from the obtained image by applying a mask that removes the pixels that do not correspond to the player's clothes. The second histogram is obtained from the image resulting from applying a mask that deletes the pixels that do not correspond to the player's skin. For this histogram the Hue channel of HSV color space is used [19]. In frame 1 the same position of the region of interest corresponding to frame 0 is taken, and for each histogram described earlier the following steps are executed [20]:

1. For each pixel value of the image the corresponding bin in the histogram is located.
2. The value associated with the selected bin is taken.
3. The value of the bin is stored into a new image.

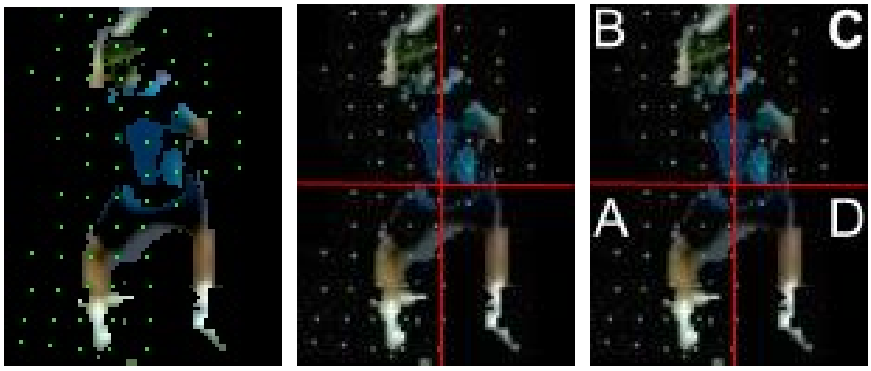
Fig. 4: Region of interest.

The values stored in each output image represent the probability that a pixel in the input image belongs to the area of interest represented by the histogram. For this case, the values stored correspond to the skin and clothing color respectively. Then, the values associated with these images are added and the result along with the region of interest of the previous frame are used as input to the Meanshift algorithm [21]. As a result, a new region of interest corresponding to the player's position in the current frame is obtained. This process is repeated for all the frames of the video.

Optical flow is then used to describe classes in a robust discriminative manner (Fig. 5a), which is calculated by the Farnebäck algorithm [6]. The displacement matrix is divided into four regions (Fig. 5b) and four attributes, denoted by h0, h1, h2, h3, are obtained. Each hi indicates the number of points in the i quadrant in which there was movement from one frame to another. This means that the variation of the optical flow is represented in each region (Fig. 5c).

These numerical values are replaced by labels A, B, C and D, depending on the number of points in each region. For instance, from h0=13, h1=16, h2=12 and h3=22 the resulting mapping is h0=C, h1=B, h2=D and h3=A.

The goal of the last step is to obtain a discretization of the optical flow data to make the attribute matching process more flexible during the construction and validation of the classifier. The decision to discretize this data was taken after empirically verifying that better results were obtained by using labels instead of numeric values.



(a) (b) (c)

Fig. 5: (a) Optical Flow, (b) Displacement matrix divided into four regions, (c) Regions ordered by the number of displacements.

Finally, CRFs are used to carry out the construction of the classifier that allows to perform the task of recognizing the two types of shots proposed in this paper. This classifier is trained with the training set obtained in the previous step.

For the training and classification stages the tool CRFSuite [22] is used. This tool has been used primarily for natural language processing.

CRFsuite has a specific data format used for training and tagging. Data consists of a set of item sequences, each one represented by consecutive lines and terminated by an empty line. An item sequence consists of a series of items whose characteristics (labels and attributes) are described in the lines. An item line begins with its label, followed by its attributes separated by TAB ('\t') characters. CRFsuite disregards the naming convention as well as the design of the labels and attributes, treating them as mere strings.

In the context of this paper, the input consists of a plain text file, where each video clip used for the training stage is represented by a sequence of lines standing for frames of the video. In turn, each line consists of 5 columns. The first corresponds to the label of the frame (i.e., the actual kind of the stroke in the frame), and the remaining four columns correspond to the optical flow representation, which was previously described in this paper.

Table 1: Data format

Frame 1
rightSwing h0=C h1=B h2=D h3=A
rightSwing h0=B h1=C h2=D h3=A
rightSwing h0=B h1=D h2=A h3=C
rightSwing h0=D h1=A h2=B h3=C
rightSwing h0=D h1=C h2=B h3=A
...
Frame 2
rightSwing h0=D h1=A h2=C h3=B
rightSwing h0=C h1=B h2=A h3=D
rightSwing h0=B h1=A h2=C h3=D
...
Frame n
rightSwing h0=C h1=B h2=A h3=D
rightSwing h0=C h1=A h2=B h3=D
...

3.2 Validation

This step consists in the classification of a video input in one of the two classes defined in the training process. The input video is processed using the tracking and extraction of features process previously explained. These features are properly formatted and used as input to the classifier, which is in charge of predicting the class associated with the input video.

4 Evaluation

The effectiveness of the proposed methods was evaluated with three measures: precision (1), recall (2) and F1 (3). The F1-Score [23] can be interpreted as a weighted average of the precision and recall, where an F1 scored reaches its best value at 1 and worst score at 0. To evaluate the model K-fold cross validation was used.

$$precision = \frac{(TruePositives)}{(TruePositives + FalsePositives)} \tag{1}$$

$$recall = \frac{(TruePositives)}{(TruePositives + FalseNegatives)} \tag{2}$$

$$F1 = \frac{2 * (precision * recall)}{(presicion + recall)} \tag{3}$$

A conventional method of calculating the performance of a classifier based on precision and recall is called macro-averaging. Macro-averaged scores are calculated by first calculating precision and recall for each class and then taking the average of these.

$$Macro\text{-}averaged\ precision_i = \frac{(PR_i + PL_i)}{2} \tag{4}$$

$$Macro\text{-}averaged\ recall_i = \frac{(RR_i + RL_i)}{2} \tag{5}$$

$$Macro\text{-}averaged\ F1_i = \frac{(FR_i + FL_i)}{2} \tag{6}$$

K-fold cross validation is a special case of a more general method called cross validation. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on the training set and validating the analysis on the testing set.

In the K-fold cross validation method, the data is divided into k subsets, and the cross-validation process is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are used as the training set. Then, the average error across all k trials is computed. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. The advantage of K-fold cross validation is that all observations are used for both training and validation, and each observation is used for validation exactly once. Another advantage is that it matters less how the data gets divided.

The classifier was evaluated using 160 video clips and 5-fold cross validation was applied. For each fold, 128 clip were used for training and 32 - 16 right-swing and 16 left-swing - were used for testing. The dataset can be downloaded from [24]. The different confusion matrices are shown below. Table 3 shows

Table 2: 5-folds cross-validation average

		Predicted Class	
		Left-swing	Right-swing
Actual class	Left-swing	**0.85**	0.15
	Right-swing	0.1625	**0.8375**

Table 3: Measures for each fold

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Precision of Right-swing Class	0.79	1	0.8	0.76	0.93
Precision of Left-swing Class	0.92	0.89	0.76	0.8	0.84
Recall of Right-swing Class	0.94	0.87	0.75	0.81	0.81
Recall of Left-swing Class	0.75	1	0.81	0.75	0.94
F1 of Right-swing Class	0.86	0.93	0.77	0.78	0.87
F1 of Left-swing Class	0.83	0.94	0.78	0.77	0.89
Macro-averaged Precision	0.86	0.95	0.78	0.78	0.89
Macro-averaged Recall	0.85	0.94	0.78	0.78	0.88
Macro-averaged F1	0.85	0.94	0.78	0.78	0.88

precision, recall, F1, Macro-averaged, Macro-recall and Macro-F1 for each class respectively.

$$Mean\ Macro\text{-}averaged\ precision = 0.85 \tag{7}$$

$$Mean\ Macro\text{-}averaged\ Recall = 0.87 \tag{8}$$

$$Mean\ Macro\text{-}averaged\ F1 = 0.85 \tag{9}$$

5 Conclusions and Future work

This article presented a novel framework for identifying different kinds of tennis shots during tennis matches. The proposed approach combines the application of video processing techniques for region of interest detection and feature extraction in tennis videos, and CRFs for action recognition. The proposed framework has been evaluated using a dataset consisting of 160 manually labeled videos. The obtained results confirm the suitability of the proposed techniques in the domain under analysis and hold great promise for action recognition in other areas.

To the best of the authors knowledge this is the first attempt to apply CRF to action recognition in tennis videos. This problem has also been addressed in [9] using support vector machines instead CRF for classification. As part of our future work we plan to compare this approach to ours. However, this comparison will be limited due to the unavailability of the datasets used in that earlier work and the inaccuracies that could result from attempting to implement our own version of the methods described in that work.
Also, further analysis will be performed in order to improve aspects such as tracking, background substraction, feature extraction for action modeling, and representation of CRFs' input data. Regarding the classification stage, a future development will consist in an implementation of a CRFs tool tailored to the image and video processing domain. This will avoid the need of adapting a natural language processing tool to the problem domain at hand.
In addition, in order to further evaluate the method, the system will be trained to recognize a greater number of tennis shots.

Acknowledgment: The authors are grateful to the anonymous reviewers, who helped in making this paper more clear and concise. José Francisco Manera and Jonathan Vainstein are supported by a Doctoral Scholarship of the CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas - Argentina). Ana Maguitmans research is partially supported by CONICET (PIP 112-200901-00863) and SGCyT-UNS (PGI 24/N029). Claudio Delrieuxs research is partially supported by a SeCyT-UNS grant (Secretaría de Ciencia y Técnica - Universidad Nacional del Sur).

References

1. Takahashi, M., Naemura, M., Fujii, M., Satoh, S.: Human action recognition in crowded surveillance video sequences by using features taken from key-point trajectories. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on. (2011) 9–16
2. Kamijo, S., Matsushita, Y., Ikeuchi, K., Sakauchi, M.: Incident detection at intersections utilizing hidden markov model. In: Proceedings of 6th World Congress on Intelligent Transport System(ITS). (November 8-12, 1999)
3. Zhu, G., Xu, C., Huang, Q.: Player action recognition in broadcast tennis video with applications to semantic analysis of sports game. In: in Proc. ACM Multimedia, 2006. (2006) 431–440
4. Lafferty, J.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, Morgan Kaufmann (2001) 282–289
5. Sutton, C., McCallum, A.: An introduction to conditional random fields. Arxiv preprint arXiv:1011.4088 (2010)
6. Farnebäck, G.: Two-frame motion estimation based on polynomial expansion. In Bigun, J., Gustavsson, T., eds.: Image Analysis. Volume 2749 of Lecture Notes in Computer Science. Springer Berlin Heidelberg (2003) 363–370
7. Miyamori, H., Iisaku, S.i.: Video annotation for content-based retrieval using human behavior analysis and domain knowledge. In: Proceedings of the Fourth IEEE

International Conference on Automatic Face and Gesture Recognition 2000. FG'00, Washington, DC, USA, IEEE Computer Society (2000) 320–
8. Efros, A.A., Berg, A.C., Mori, G., Malik, J.: Recognizing action at a distance. In: Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2. ICCV'03, Washington, DC, USA, IEEE Computer Society (2003) 726–
9. Zhu, G., Xu, C., Gao, W., Huang, Q.: Action recognition in broadcast tennis video using optical flow and support vector machine. In: Proceedings of the 2006 international conference on Computer Vision in Human-Computer Interaction. ECCV'06, Berlin, Heidelberg, Springer-Verlag (2006) 89–98
10. Sha, F., Pereira, F.: Shallow parsing with conditional random fields. In: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1. NAACL '03, Stroudsburg, PA, USA, Association for Computational Linguistics (2003) 134–141
11. Sutton, C., McCallum, A., Rohanimanesh, K.: Dynamic conditional random fields: Factorized probabilistic models for labeling and segmenting sequence data. Journal of Machine Learning Research **8** (March 2007) 693–723
12. McCallum, A.: Efficiently inducing features of conditional random fields. In: Proceedings of the Nineteenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-03), San Francisco, CA, Morgan Kaufmann (2003) 403–410
13. Culotta, A., Bekkerman, R., Mccallum, A.: Extracting social networks and contact information from email and the web. In: In Proceedings of CEAS-1. (2004)
14. Sato, K., Sakakibara, Y.: RNA secondary structural alignment with conditional random fields. In: ECCB/JBI. (2005) 242
15. Yan Liu, Jaime Carbonell, P.W., Gopalakr-ishnan, V.: Protein fold recognition using segmentation conditional random fields (scrfs). Journal of Computational Biology, 13 (2) 394-406 (2006)
16. He, X., Zemel, R.S., Carreira-Perpiñán, M.A.: Multiscale conditional random fields for image labeling. In: Proceedings of the 2004 IEEE computer society conference on Computer vision and pattern recognition. CVPR'04, Washington, DC, USA, IEEE Computer Society (2004) 695–703
17. Quattoni, A., Collins, M., Darrell, T.: Conditional random fields for object recognition. In: In NIPS, MIT Press (2004) 1097–1104
18. Jain, N., Chaudhury, S., Roy, S.D., Mukherjee, P., Seal, K., Talluri, K.: A novel learning-based framework for detecting interesting events in soccer videos. In: Proceedings of the 2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing. ICVGIP'08, Washington, DC, USA, IEEE Computer Society (2008) 119–125
19. Zarit, B.D., Super, B.J., Quek, F.K.H.: Comparison of five color models in skin pixel classification. In: In ICCV'99 Int l Workshop on. (1999) 58–63
20. OpenCV: Opencv: Backprojection. `http://docs.opencv.org/doc/tutorials/imgproc/histograms/back_projection/back_projection.html` ((accessed March 9, 2013))
21. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. IEEE Trans. Pattern Anal. Mach. Intell. **25**(5) (may 2003) 564–575
22. Okazaki, N.: Crfsuite: a fast implementation of conditional random fields (crfs) (2007)
23. Wikipedia: Presicion, recall and f1. `http://en.wikipedia.org/wiki/Precision_and_recall` ((accessed March 9, 2013))

24. Imaglabs: Imaglabs tennis video dataset. http://members.imaglabs.org/jonathan.vainstein/tennis_dataset.tar.gz