ADVANCED PLAYER ACTIVITY RECOGNITION BY INTEGRATING BODY POSTURE AND MOTION INFORMATION

Marco Leo, Tiziana D'Orazio, Paolo Spagnolo and Pier Luigi Mazzeo
Institute of Intelligent Systems for Automation, Italian National Council Research, via Amendola 122/d, Bari , Italy

Keywords: Human pose estimation, Contourlet transform, Neural networks, Soccer player activity recognition.

Abstract: Human action recognition is an important research area in the field of computer vision having a great number of real-world applications. This paper presents a multi-view action recognition framework able to extract human silhouette clues from different synchronized static cameras and then to validate them introducing advanced reasonings about scene dynamics. Two different algorithmic procedures have been introduced: the first one performs, in each acquired image, the neural recognition of the human body configuration by using a novel mathematic tool named Contourlet transform. The second procedure performs, instead, 3D ball and player motion analysis. The outcomes of both procedures are then properly merged to accomplish the final player activity recognition task. Experimental results were carried out on several image sequences acquired during some matches of the Italian Serie A soccer championship.

1 INTRODUCTION

Human action recognition aims at automatically ascertaining the activity of a person, i.e. to identify if someone is walking, dancing, or performing other types of activities. It is an important area of research in the field of computer vision and the ever growing interest in it is fueled, in part, by the great number of real-world applications such as surveillance scenarios, content-based image retrieval, human-robot interaction, sport video analysis,smart rooms etc.

Human action recognition has been a widely studied topic and extensive reviews can be found in (Weiming et al., 2004) and (Agarwal and Triggs, 2006). Human activity recognition approaches are categorized on the basis of the representation of the human body: representation can be extracted either from a still image or a dynamic video sequence.

In general, performing human action recognition from video sequences requires complex models for understanding the dynamics (Gorelick et al., 2007), (Jhuang et al., 2007), (Niebles et al., 2008), (Liu et al., 2008). On the other side recent studies demonstrated that static human pose encapsulates many useful clues for recognizing the ongoing activity (Ikizler and Duygulu, 2007), (Goldenberg et al., 2005), (Lu and Little, 2006), (Zhang et al., 2007), (Thurau, 2007).

Unfortunately, due to possible large variations in body appearance, both static and dynamic representations have a considerable failure rate unless specific databases are used.

In this paper we introduce a new multi-view action recognition system that extracts human silhouette clues from different synchronized static cameras and then it validates them by introducing some reasonings about scene dynamics. The experimental results were carried out on several image sequences acquired during some matches of the Italian "Serie A" soccer championship.

2 SYSTEM OVERVIEW

Six high resolution cameras (labeled as FGi, where i indicates the $i-th$ camera) have been placed on the two sides of the pitch assuring double coverage of almost all the areas by either adjacent or opposite cameras. In figure 1 the location of the cameras is shown. The acquired images are transferred to six processing nodes by fiber optic cables. The acquisition process is guided by a central trigger generator that guarantees synchronized acquisition between all the cameras. Each node, using two hyper-threading processors, records all the images of the match on its internal

Leo M., D'Orazio T., Spagnolo P. and Luigi Mazzeo P. (2009).
ADVANCED PLAYER ACTIVITY RECOGNITION BY INTEGRATING BODY POSTURE AND MOTION INFORMATION.
In *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, pages 261-266
DOI: 10.5220/0001754002610266

storage unit, displays the acquired images and, simultaneously, processes them with parallel threads, in an asynchronous way with respect to the other nodes.

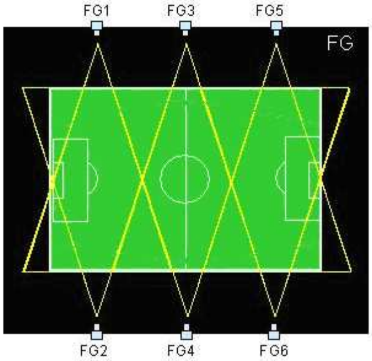

Figure 1: The location of the cameras around the pitch.

The six processing nodes, are connected to a central node, having the supervisor function. It synchronizes data coming from nodes and performs high level processing.

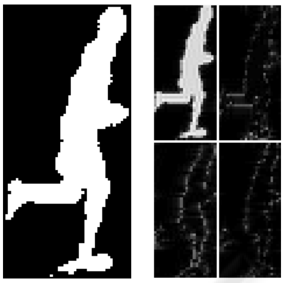

Figure 2: The binarized player silhouette (on the left) and its Contourlet representation (on the right).

Each node uses a background subtraction algorithm for motion detection. It is based on a modified version of a well known approach for background creation and maintenance (Kanade et al., 1998). Information relative to moving objects is then sent to two parallel processing threads: the first one performs human blob detection, classification (Spagnolo et al., 2007) and tracking (D'Orazio et al., 2007) as well as neural player activity recognition using static representation by Contourlet transform; the second one performs ball detection by means of a correlation based approach using six reference sets containing some ball examples acquired in different positions with respect to the camera (near, far, very far) and in different lighting conditions (sunny days, evening or cloudy days).

At each trigger pulse, the outcomes of the algorithmic procedures running on each processing node, are sent to the central node that analyzes them in order to localize the ball and the players on a virtual play-field, to compute their trajectories and to validate player activity by using motion information.

3 PLAYER BODY POSTURE ESTIMATION

The first step in the proposed framework deals with the recognition of the player body configuration *(postures)* in each of the different cameras. For this purpose, in this paper, a learning based approach is used: first of all the player silhouettes are binarized, mirrored (only those coming from the three cameras FG2, FG4 and FG6) in order to use the same left-right labeling system for body configuration, re-sized to avoid scaling effects, and described by using Contourlet coefficients that are extracted via a double iterated filter bank structure providing a flexible multi-resolution, local and directional image expansion (Do and Vetterli, 2005).

The new Contourlet representation is then provided as input to a back propagation neural network able to recognize seven different human configurations associated to seven player activities: walking, running left, running right, running front-back, still, shooting left, shooting right.

The neural network architecture (experimentally set) consists of three processing layers: 30 hidden neurons with sigmoidal activation functions and 6 output neurons with *softmax* activation functions (Bishop, 1995) are used. The neural output values are managed as follows: the greatest output value is considered and if it is greater than $th = 0,5$ the input patch is labeled by the corresponding activity, otherwise it is labeled as undetermined. In figure 2, on the left the initial binary silhouette of a running player is reported whereas on the right the relative Contourlet representation is shown.

4 BALL AND PLAYER TRAJECTORY COMPUTATION

The second algorithmic procedure runs on the processing unit with supervisor function. The supervisor makes use of a virtual play-field (having the same dimension of the real play field) to project extracted information: in particular the player and referee data are projected onto the virtual play-field by homographic transformation assuming that their feet are always in contact with the ground. Anyway, the projection of the same player using data relative to different cameras are not coincident due to the different segmentations into the image planes caused by different appearances of the same player (different position with respect the camera, different lighting conditions, shadows and so on). To overcome this drawback, the mid-point of the line connecting the different projections of the player in the virtual play-field has been considered for further processing.

The projection of the ball position in to the virtual play-field requires, instead, a different procedure considering that the ball is not always in contact with the ground. The 3D ball position has to be then firstly recovered by triangulation (if ball information coming from two opposite or adjacent views is available) and then its projection onto the virtual play field can be performed.

In figure 3 the virtual play-field is reported. The red and cyan rectangles indicate the player positions computed by merging data coming from two opposite or adjacent views (relative IDs assigned from nodes are also reported) whereas the ball position is indicated by the yellow cross. The white lines behind each object indicate recent ball and players displacements.

Starting from the estimated ball and players positions their temporal trajectories can be computed. The player trajectories in the virtual play-field cannot be mathematically modeled by straight lines or curves; they vary continuously in an unpredictable way and then they can be represented only collecting the player positions into the play-field.

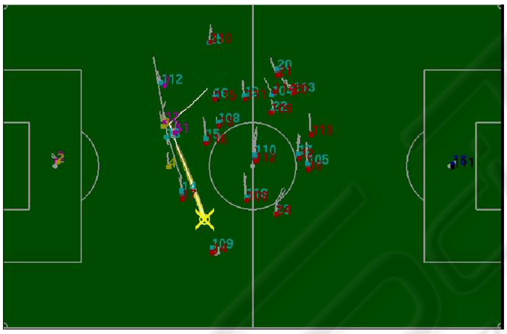

Figure 3: The virtual play-field.

For the ball, instead, trajectories in the virtual play-field can be approximated by straight lines: this allows the system to predict the successive positions, to recover missed intermediate ones, and to introduce high level reasonings useful to understand the soccer game developing. For the sake of precision we have to explain that we dispose of both 3D and 2D ball trajectories. In order to detect shots, as abrupt changes of trajectories, we consider in this paper only 2D trajectories, obtained by projecting the 3D ball positions in to the virtual play-field. This simplification allows the system to avoid false shot detections when there are ball rebounds on the field.

5 MULTI-VIEW PLAYER ACTIVITY RECOGNITION

The outcomes of the algorithmic procedures described in the sections 3 and 4 are, finally, given as input to a higher level functional step running on the supervisor unit that performs a multi-view player activity recognition. To do that, first of all, the supervisor processing unit merges body posture information coming from different views of the same player: the M available estimation scores extracted by the single view procedure (one for each camera acquiring the considered player) for the $i-th$ player are averaged to obtain k values of MPV (*M*ulti-view *P*robability *V*alue) :

$$(MPV)_k = P(X_k|z_1...z_M) = \frac{1}{n}\sum_{1}^{M} p(X_k|z_j) \quad k = 1,...N$$

where X is the player posture class, z_i are the single view estimated configurations and N indicates the maximum number of body configurations classes to be recognized. In this way a global estimation score is estimated for each of the k configuration classes. The problem becomes now how to decide which body configuration class has to be associated to the $i-th$ on the basis of the relative available P_k with $k = 1,...N$.

To solve this problem a preliminary statistical evaluation of the neural outcomes is done: the values relative to both correct and incorrect occurrences during a preliminary experimental phase are considered and they are then used to estimate the relative gaussian probability distributions. This demonstrates that the most probable values in case of correct body configuration estimations in a single view are close to 0.85, whereas in case of wrong estimations they are close to 0.5.

Starting from this statistical consideration a multi-view decision rule, based on available MPVs, is introduced: the player's body configuration K is associated to the *i-th* player if

$$\{(K = \arg\max_{k=1,...6}(MPV)_k) \wedge ((MPV)_K > th) \wedge$$

$$\wedge((MPV)_i < thi = \{1...N, i \neq K\})$$

where th is the intersection point of the above estimated $pdfs$. If these conditions are not simultaneously satisfied the considered player body configuration is labeled as undetermined.

5.1 Integration of Motion Information

After that, a static multi-view player activity estimation is available for each player in the scene and it can be validated by using motion information. In particular ‘running’, ‘walking’ and ‘still’ activities are validated by using motion information of the relative player, whereas ’shooting’ activity is validated taking under consideration also the 3D ball trajectory. In fact for an estimated ’still’ player the system checks his motion: if, considering the last three frames, his position in the virtual playing field does not significatively change, the player activity is definitively labeled as ‘still’. The same approach, based on the analysis of the position changes into the play-field, is also used for validating estimated ‘running’ and ‘walking’ players: in this case the estimated player body posture is validated if player position (considering the last three frames) changes according to the common running and walking velocity values for a human being whereas running directions (left, right, front-back) are validated considering the recovered player trajectory into the play-field. In particular a player is validated as ’walking’ if this velocity in the virtual play-field varies between 3 km/h and 6 km/h; running activity is instead validated if the velocity of the relative player is greater than 6 km/h.

A quite different approach is finally used for validating estimated ’shooting’ players: in this case both ball and player motion information are used. In fact, ’shooting’ player is validated if : 1) the 3D ball trajectory indicates that the ball is really going away from the considered player (as expected for the kicking player) 2) the player is near (at lest 2 meters) to $P(x_s, y_s)$, i.e. the intersection point of two consecutive validated ball trajectories.

In this way motion information helps to improve player activity recognition but, at the same time, could solve some of the previously unclassified activity occurrences: in fact motion reasonings can be separately applied to the different incoherent human body configuration outcomes in order to point out the most likely one.

Finally notice that in the case of incoherence between estimated player body configurations and motion information the relative player activity is definitively labeled as ’undetermined’.

6 EXPERIMENTAL RESULTS

The proposed multi-steps method was applied to several image sequences acquired during some matches of the Italian “Serie A” championship.

Experiments were carried out on a set of 3500 different pairs of synchronized patches (7000 total patches) relative to players acquired during different soccer matches. These patches were preliminarily labeled by a human operator who assigned to each pair one of the seven considered activities: ***A****-Running right side;* ***B****-Running left side;* ***C****-Running front-back;* ***D****-Walking;* ***E****-Still;****F****- Shooting left side;* ***G****-Shooting right side.*

The ground truth relative to the 7000 patches is reported in table 1.

Table 1: The ground truth relative to the 7000 considered binary patches used in the experimental phase.

.

A	B	C	D	E	F	G
1680	**1844**	**412**	**1410**	**756**	**476**	**422**

The first step of the experimental phase concentrates on the recognition of the human activity on each single image by using Contourlet representation and a neural classifier as described in section 3.

The results of this first experiment are reported in table 2. Because of lack of space the first row reports the capital letters relative to each of the seven considered body configurations as in the above list.

Table 2: The scatter matrix relative to the first experiment regarding the recognition of the player activity recognition by single image.

A	B	C	D	E	F	G	und.
1243	9	218	160	0	0	0	50
0	**1624**	0	111	108	0	0	1
0	0	**305**	33	25	29	0	20
0	0	14	**1269**	30	56	0	41
0	21	62	0	**620**	0	0	53
9	0	16	0	0	**432**	0	19
13	5	5	0	0	0	**387**	12

The experimental results reported in table 2 were very encouraging: almost 84% of the testing patches were automatically labeled by the system in the same way as the human operator. Some miss-classifications happened due to the similarity of appearance, under certain conditions, of the body silhouettes relative to players performing different activities. For example, in figure 4, three wrongly classified patches are reported: the player on the left was classified as running towards the camera by the human operator whereas the automatic system consider him as running right. The player in the center was instead classified as running towards the camera by the human operator whereas the automatic system classified him as still. Finally, the player on the right was labeled as kicking by the human operator and running right by the automatic system. As you can see in this cases it is not easy to definitively decide real player configurations and then you have to consider that experimental results strongly depend, on the operator that generated the ground truth.

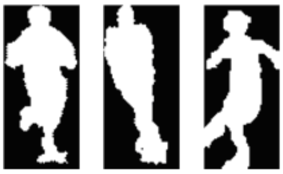

Figure 4: Three different cases where the proposed system missrecognized player body configurations.

The activity data coming from this first experimental step were then merged, for each pair of opposite cameras, by using the procedure described in section 5. In table 3 the multi-view activity estimation results on the 3500 pairs of binary patches are reported.

Table 3: Activity recognition performance integrating information coming from different camera views. The test set consists of 3500 pairs of binary patches.

A	B	C	D	E	F	G	und.
822	2	3	2	0	0	5	6
0	**890**	0	12	9	6	0	5
0	0	**186**	2	8	7	0	3
0	0	10	**666**	7	8	0	14
0	9	5	0	**345**	0	0	19
0	0	3	0	0	**227**	0	8
5	0	3	0	0	0	**199**	4

More than 95% of the 3500 tested pairs were correctly recognized. Less than 2% of the tested patches were not classified due to the ambiguities in the probability values provided by the neural algorithms running on the images coming from each camera.

In figure 5, the two pictures on the left report the player's silhouettes acquired by two opposite cameras at a shot instant. In this case the analysis of human body configuration performed on each view agreed and they indicated that the player was shooting the ball with probability values respectively of 0.97 and 0.89. The two pictures on the right show, instead, a case in which the procedure running on each single view disagreed: one camera recognized the player as walking (probability value 0.87) and the other one as still (probability value 0.54). The multi-view approach solved this ambiguity and labeled the player walking as the human operator did. Finally, the remaining 3% of the tested patches were miss-classified.

Finally the validation procedure based on motion information described in section 5.1 was tested in order to verify its capability to improve player activity recognition. The final results of the proposed player activity recognition approach are then reported in table 4.

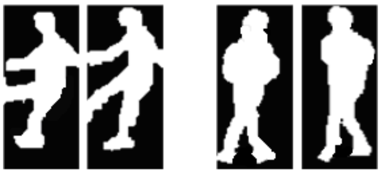

Figure 5: Different pairs of patches containing the same player acquired from different cameras.

Table 4: Activity recognition performance after validation by motion information.

A	B	C	D	E	F	G	und.
833	0	0	0	0	0	1	6
0	**912**	0	3	1	0	0	6
0	0	**196**	0	3	0	0	7
0	0	4	**689**	2	0	0	10
0	3	1	0	**357**	0	0	17
0	0	0	0	0	**230**	0	8
0	0	0	0	0	0	**204**	7

Introducing activity validation by motion information drastically reduces both uncertainty occurrences and increases correct classification. In figure 6 two examples pointing out the benefits of using motion information are reported. On the left two patches (acquired from FG3 and FG4) are relative to a player kicking the ball. The neural approach did not classify player activity because, for the patch on the left, it rightly recognizes the player as shooting left but, for the patch on the right, it erroneously classifies the player as running left. Introducing reasonings about motion and ball proximity the system verified that both the ball was close to the player and the distance of the player from the intersection point of the ball trajectory was very small. For these reasons the player was correctly classified as shooting the ball. On the right, instead, the two patches are relative to a player running right (acquired from FG5 and FG6). Unfortunately both patches were classified as ”walking” by the neural approach based on Contourlet representation (most probably due to perspective distortions). The motion validation procedure did not validate the player as walking due to his velocity on the pitch (9 km/h) and then his activity was considered as ’undermined’ avoiding a wrong classification.

Finally in figure 7 an example in which motion information did not solve miss-classification is reported. The two players were classified as shooting right by the multi-view approach described in section 5. Unfortunately the ball was very close to the play-

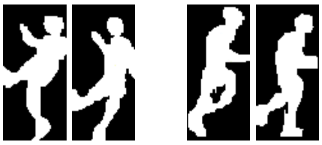

Figure 6: Two examples in which motion information overcome drawback of the neural approach for player activity recognition.

ers and it was going away from them and then both of them were also validated as shooting right by the procedure based on motion information integration. The ground truth, instead, indicated that the ball was kicked by the player having a white strip and that the player with blue strip is instead just running right side.

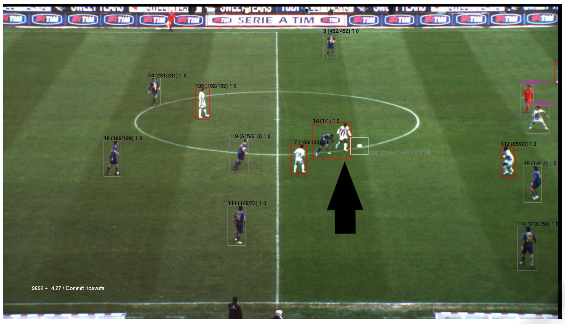

Figure 7: An example in which motion information did not solve activity miss-classification.

REFERENCES

Agarwal, A. and Triggs, B. (2006). Recovering 3d human pose from monocular images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(1):44–58.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Do, M. N. and Vetterli, M. (2005). The contourlet transform: An efficient directional multiresolution image representation. *IEEE Transactions on Image Processing*, 14(12):2091–2106.

D'Orazio, T., Leo, M., Spagnolo, P., Mazzeo, P., Mosca, N., and Nitti, M. (2007). A visual tracking algorithm for real time people detection. In *Wiamis2007: Proceedings of the International Workshop on Image Analysis for Multimedia Interactive Services*.

Goldenberg, R., Kimmel, R., Rivlin, E., and Rudzsky, M. (2005). Behavior classification by eigendecomposition of periodic motions. *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews*, 38(7).

Gorelick, L., Blank, M., Shechtman, E., Irani, M., and Basri, R. (2007). Actions as space-time shapes. *Transactions on Pattern Analysis and Machine Intelligence*, 29(12):2247–2253.

Ikizler, N. and Duygulu, P. (2007). Human action recognition using distribution of oriented rectangular patches. In *HUMO07*, pages 271–284.

Jhuang, H., Serre, T., Wolf, L., and Poggio, T. (2007). A biologically inspired system for action recognition. In *ICCV07*, pages 1–8.

Kanade, T., Collins, R., Lipton, A., Burt, P., and Wixson, L. (1998). Advances in cooperative multi-sensor video surveillance. In *Proceedings of DARPA Image Understanding Workshop, volume 1, pages 3–24, November*.

Liu, J., Ali, S., and Shah, M. (2008). Recognizing human actions using multiple features. In *International Conference on Computer Vision and Pattern Recognition CVPR08*.

Lu, W. and Little, J. (2006). Simultaneous tracking and action recognition using the pca-hog descriptor. In *CRV06*, pages 6–6.

Niebles, C., Wang, H., and Fei-Fei, L. (2008). Unsupervised learning of human action categories using spatial-temporal words. *International Journal of Computer Vision*, in press:2247–2253.

Spagnolo, P., Mosca, N., Nitti, M., and Distante, A. (2007). An unsupervised approach for segmentation and clustering of soccer players. In *IMVIP2007: Proceedings of the International Machine Vision and Image Processing Conference*.

Thurau, C. (2007). Behavior histograms for action recognition and human detection. In *HUMO07*, pages 299–312.

Weiming, H., Tieniu, T., Liang, W., and Steve, M. (2004). A survey on visual surveillance of object motion and behaviors. *IEEE transactions on systems, man and cybernetics. Part C, Applications and reviews*, 34(3).

Zhang, L., Wu, B., and Nevatia, R. (2007). Detection and tracking of multiple humans with extensive pose articulation. In *ICCV07*, pages 1–8.