

Detection and recognition of sports(wo)men from multiple views

Damien Delannay, Nicolas Danhier, and Christophe De Vleeschouwer
Université catholique de Louvain, Belgium.

Abstract—The methods presented in this paper aim at detecting and recognizing players on a sport-field, based on a distributed set of loosely synchronized cameras. Detection assumes player verticality, and sums the cumulative projection of the multiple views’ foreground activity masks on a set of planes that are parallel to the ground plane. After summation, large projection values indicate the position of the player on the ground plane. This position is used as an anchor for the player bounding box projected in each one of the views. Within this bounding box, the regions provided by mean-shift segmentation are sorted out based on contextual features, e.g. relative size and position, to select the ones that are likely to correspond to a digit. Normalization and classification of the selected regions then provides the number and identity of the player. Since the player number can only be read when it faces towards the camera, graph-based tracking is considered to propagate the identity of a player along its trajectory.

I. INTRODUCTION

In today’s society, content production and content consumption are confronted with a fundamental mutation. Two complementary trends are observed. On the one hand, individuals become more and more heterogeneous in the way they access the content. They want to access dedicated content through a personalized service, able to provide what they are interested in, when they want it and through the communication channel of their choice. On the other hand, individuals and organizations get easier access to the technical support and facilities required to be involved in the content creation and diffusion process.

In this paper, we describe video analysis tools that participate to the future evolutions of the content production industry towards automated infrastructures allowing content to be produced, stored, and accessed at low cost and in a personalized and dedicated manner. More specifically, our targeted application considers the autonomous and personalized summarization of sport events, without the need for costly handmade processes. In the application scenario supported by the provided dataset, the acquisition sensors cover a basket-ball court.

Part of this work has been funded by the FP7 European project APIDIS, and by the Belgian NSF.

Distributed analysis and interpretation of the scene is then exploited to decide what to show about an event, and how to show it, so as to produce a video composed of a valuable subset from the streams provided by each individual camera. In particular, the position of the players provides the required input to drive the autonomous selection of viewpoint parameters[5], whilst identification and tracking of the detected players supports personalization of the summary, e.g. through highlight and/or replay of preferred player’s actions[4].

II. SYSTEM OVERVIEW

To demonstrate the concept of autonomous and personalized production, the European FP7 APIDIS research project (www.apidis.org) has deployed a multi-camera acquisition system around a basket-ball court. The acquisition setting consists in a set of 7 calibrated IP cameras, each one collecting 2 Mpixels frames at a rate higher than 20 frames/sec. After an approximate temporal synchronization of the video streams, this paper investigates how to augment the video dataset based on the detection, tracking, and recognition of players along the game.

Figure 1 surveys our proposed approach to compute and label players tracks. After joint multiview detection of people standing on the ground field at each time instant, a graph-based tracking algorithm matches positions that are sufficiently close -in position and appearance- between successive frames, thereby defining a set of potentially interrupted disjoint tracks, also named partial tracks. In parallel, as depicted in Figure 5, image analysis and classification is considered for each frame of each view, to recognize the digits that potentially appear on the shirts of detected objects. This information is then aggregated over time to label the partial tracks.

The major contributions of this paper have to be found in the proposed people detection solution, which is depicted in Figure 2. In short, the detection process follows a bottom-up approach to extract denser clusters in a ground plane occupancy map that is computed based on the projection of foreground activity masks. Two

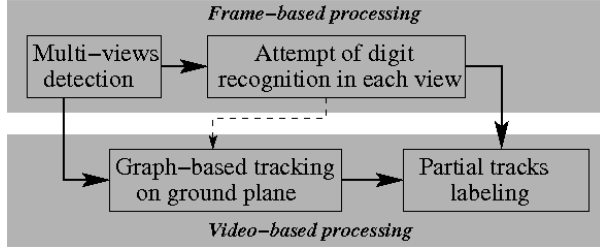


Fig. 1. Players tracks computation and labeling pipeline. The dashed arrow reflects the optional inclusion of the digit recognition results within the appearance model considered for tracking.

fundamental improvements are proposed compared to the state-of-the art. First, the foreground activity mask is not only projected on the ground plane, as recommended in [9], but on a set of planes that are parallel to the ground. Second, an original heuristic is implemented to handle occlusions, and alleviate the false detections occurring at the intersection of the masks projected from distinct players' silhouettes by distinct views. Our simulations demonstrate that those two contributions quite significantly improve the detection performance.

The rest of the paper is organized as follows. Sections III, V, and IV respectively focus on the detection, tracking, and recognition problems. Experimental results are presented in Section VI to validate our approach. Section VII concludes.

III. MULTI-VIEW PEOPLE DETECTION

Keeping track of people who occlude each other using a set of C widely spaced, calibrated, stationary, and (loosely) synchronized cameras is an important question because this kind of setup is common to applications ranging from (sport) event reporting to surveillance in public space. In this section, we consider a change detection approach to infer the position of players on the ground field, at each time instant.

A. Related work

Detection of people from the foreground activity masks computed in multiple views has been investigated in details in the past few years. We differentiate two classes of approaches.

On the one hand, the authors in [9], [10] adopt a **bottom-up** approach, and project the points of the foreground likelihood (background subtracted silhouettes) of each view to the ground plane. Specifically, the change probability maps computed in each view are warped to the ground plane based on homographies that have been inferred off-line. The projected maps are then multiplied together and thresholded to define the patches of the

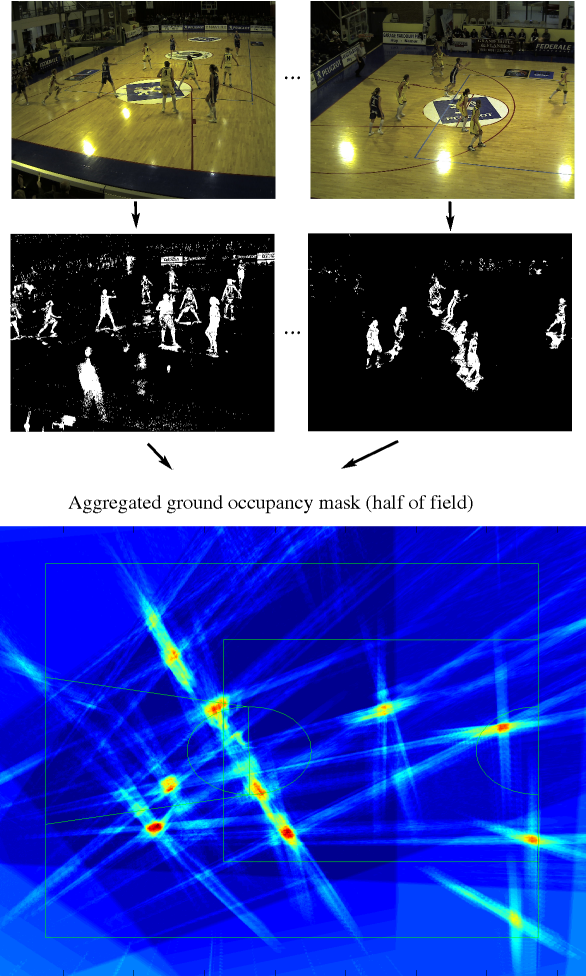


Fig. 2. Multi-view people detection. Foreground masks are projected on a set of planes that are parallel to the ground plane to define a ground plane occupancy map, from which players' position is directly inferred.

ground plane for which the appearance has changed compared to the background model and according to the single-view change detection algorithm.

On the other hand, the works in [2], [7], [1] adopt a **top-down** approach. They consider a grid of points on the ground plane, and estimate the probabilities of occupancy of each point in the grid based on the back-projection of some kind of generative model in each one of the calibrated multiple views. Hence, they all start from the ground plane, and validate occupancy hypothesis based on associated appearance model in each one of the views. The approaches proposed in this second category mainly differ based on the kind of generative model they consider (rectangle or learned dictionary), and on the way they decide about occupancy in each

point of the grid (combination of multiple view-based classifiers in [2], probabilistic occupancy grid inferred from background subtraction masks in [7], and sparsity constrained binary occupancy map for [1]).

The first category of methods has the advantage to be computationally efficient, since the decision about ground plane occupancy is directly taken from the observation of the projection of the change detection masks of the different views. In contrast, the complexity of the second category of algorithms depends on the number of ground plane points to be investigated (chosen to limit the area to be monitored), and on the computational load associated to the validation of each occupancy hypothesis. This validation process generally involves back-projection of a 3D-world template in each one of the views. With that respect, we note that, due to lens and projection distortions, even the warping of simple 3D rectangular template generally results in non-rectangular patterns in each one of the views, thereby preventing the use of computationally efficient integral images techniques. Hence, in most practical cases, the second kind of approach is significantly more complex than the first one. In return, it offers increased performance since not only the feet, but the entire object silhouette is considered to make a decision.

Our approach is an attempt to take the best out of both categories. It proposes a computationally efficient bottom-up approach that is able to exploit the entire a priori knowledge we have about the object silhouette. Specifically, the bottom-up computation of the ground occupancy mask described in Section III-B exploits the fact that the basis of the silhouette lies on the ground plane (similarly to previous bottom-up solutions), but also that the silhouette is a roughly rectangular vertical shape (which was previously reserved to top-down approaches). As a second contribution, Section III-C proposes a simple greedy heuristic to resolve the interference occurring between the silhouettes projected from distinct views by distinct objects. Our experimental results reveal that this interference was the source of many false detections while inferring the actual objects positions from the ground occupancy mask. Until now, this phenomenon had only been taken into account by the top-down approach described in [7], through a complex iterative approximation of the joint posterior probabilities of occupancy. In contrast, whilst approximate, our approach appears to be both efficient and effective.

B. Proposed approach: ground plane occupancy mask computation

Similar to [9], [10], [7], [1], our approach carries out single-view change detection independently on each

view to compute a change probability map. To this purpose, a conventional background subtraction algorithm based on mixture of gaussians modeling is implemented. To fusion the resulting binary foreground silhouettes, our method projects them to build a ground occupancy mask. However, in contrast to previous bottom-up approaches [9], [10], we do not consider projection on the ground plane only, but on a set of planes that are parallel to the ground plane, and cut the object to detect at different heights. Under the assumption that the object of interest stands roughly vertically, the cumulative projection of all those projections on a virtual top view plane actually reflects ground plane occupancy. This section explains how the mask associated to each view is computed. The next section investigates how to merge the information provided by the multiple views to detect people.

Formally, the computation of the ground occupancy mask G_i associated to the i^{th} view is described as follows. At a given time, the i^{th} view is the source of a binary background subtracted silhouette image $B_i \in \{0, 1\}^{M_i}$, where M_i is the number of pixels of camera i , $1 \leq i \leq C$. As explained above, B_i is projected on a set of L reference planes that are defined to be parallel to the ground plane, at regular height intervals, and up to the typical height of a player. Hence, for each view i , we define G_i^j to be the projection of the i^{th} binary mask on the j^{th} plane. G_i^j is computed by applying the homography warping each pixel from camera i to its corresponding position on the j^{th} reference plane, with $0 \leq j < L$. By construction, points from B_i that are labeled to 1 because of the presence of a player in the j^{th} reference plane project to corresponding top view position in G_i^j . Hence, the summation G_i of the projections obtained at different heights and from different views is expected to highlight top view positions of vertically standing players. Moreover, the larger L the more effective the method.

As L increases, the computation of G_i in a ground position $\mathbf{x}$ tends towards the integration of the projection of B_i on a vertical segment anchored in $\mathbf{x}$. This integration can equivalently be computed in B_i , along the back-projection of the vertical segment. To further speed up the computations, we observe that, through appropriate transformation of B_i , it is possible to shape the back-projected integration domains so that they correspond to segments of vertical lines in the transformed view, thereby making the computation of integrals particularly efficient through the principle of integral images. Figure 3 illustrates that specific transformation for one particular view. The transformation has been designed to

address a double objective. First, points of the 3D space located on the same vertical line have to be projected on the same column in the transformed view (vertical vanishing point at infinity). Second, vertical objects that stand on the ground and whose feet are projected on the same horizontal line of the transformed view have to keep same projected heights ratios.

Once the first property is met, the 3D points belonging to the vertical line standing above a given point from the ground plane simply project on the column of the transformed view that stands above the projection of the 3D ground plane point. Hence, $G_i(\mathbf{x})$ is simply computed as the integral of the transformed view over this vertical back-projected segment. Preservation of height along the lines of the transformed view even further simplifies computations.

For side views, these two properties can be achieved by virtually moving -through homography transforms- the camera viewing direction (principal axis) so as to bring the vertical vanishing point at infinity and ensure horizon line is horizontal. For top views, the principal axis is set perpendicular to the ground and a polar mapping is performed to achieve the same properties. Note that in some geometrical configurations, these transformations can induce severe skewing of the views.

C. Proposed approach: people detection from ground occupancy

Given the ground occupancy masks G_i for all views, we now explain how to infer the position of the people standing on the ground. A priori, we know that (i) each player induces a dense cluster on the sum of ground occupancy masks, and (ii) the number of people to detect is equal to a known value K , e.g. $K = 12$ for basket-ball (players + referees).

For this reason, in each ground location $\mathbf{x}$, we consider the sum of all projections -normalized by the number of views that actually cover $\mathbf{x}$ -, and look for the higher intensity spots in this aggregated ground occupancy mask (see Figure 2 for an example of aggregated ground occupancy mask). To locate those spots, we have first considered a **naive greedy approach** that is equivalent to an iterative matching pursuit procedure. At each step the matching pursuit process maximizes the inner product between a translated Gaussian kernel, and the aggregated ground occupancy mask. The position of the kernel which induces the larger inner-product defines the player position. Before running the next iteration, the contribution of the Gaussian kernel is subtracted from the aggregated mask to produce a residual mask. The process iterates until sufficient players have been found.

This approach is simple, but suffers from many false detections at the intersection of the projections of distinct players silhouettes from different views. This is due to the fact that occlusions induce non-linearities¹ in the definition of the ground occupancy mask. Hence, once some people are known to be present on the ground field affect the information that can be retrieved from the binary change masks in each views. In particular, if the vertical line associated to a position $\mathbf{x}$ is occluded by/occludes another player whose presence is very likely, this particular view should not be exploited to decide whether there is a player in $\mathbf{x}$ or not.

For this reason, we propose to refine our naive approach as follows.

To initialize the process, we define $G_i^0(\mathbf{x})$ to be the ground occupancy mask G_i associated to the i^{th} view (see Section III-B), and set $w_i^0(\mathbf{x})$ to 1 when $\mathbf{x}$ is covered by the i^{th} view, and to 0 otherwise. Each iteration is then run in two steps. At iteration n , the first step searches for the most likely position of the n^{th} player, knowing the position of the $(n - 1)$ players located in previous iterations. The second step updates the ground occupancy masks of all views to remove the contribution of the newly located player.

Formally, the first step of iteration n aggregates the ground occupancy mask from all views, and then searches for the denser cluster in this mask. Hence, it computes the aggregated mask G^n at iteration n as

$$G^n(\mathbf{x}) = \frac{\sum_{i=1}^C w_i^n(\mathbf{x}) G_i^n(\mathbf{x})}{\sum_{i=1}^C w_i^n(\mathbf{x})}, \quad (1)$$

and then defines the most likely position $\mathbf{x}_n$ for the n^{th} player by

$$\mathbf{x}_n = \underset{\mathbf{y}}{\operatorname{argmax}} \langle G^n(\mathbf{x}), k(\mathbf{y}) \rangle, \quad (2)$$

where $k(\mathbf{y})$ denotes a Gaussian kernel centered in $\mathbf{y}$ and whose spatial support corresponds to the typical width of a player.

In the second step, the ground occupancy mask of each view is updated to account for the presence of the n^{th} player. In the ground position $\mathbf{x}$, we consider that typical support of a player silhouette in view i is a rectangular box of width W and height H , and observe that the part of the silhouette that occludes or is occluded by the newly detected player does not bring any information about the potential presence of a player in position $\mathbf{x}$. Let $\alpha_i(\mathbf{x}, \mathbf{x}_n)$ denote the fraction of the silhouette in

¹In other words, the ground occupancy mask of a group of players is not equal to the sum of ground occupancy masks projected by each individual player.

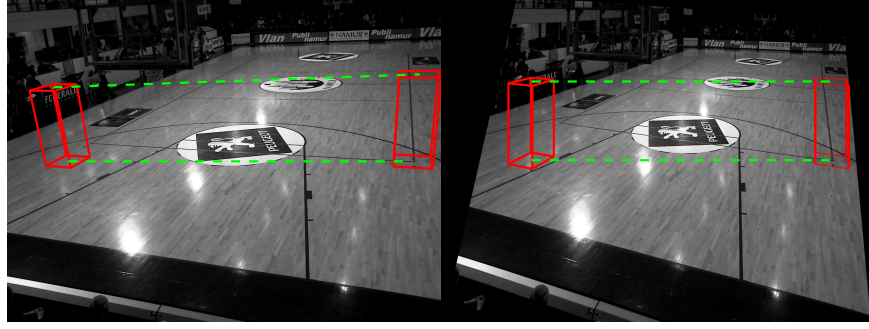


Fig. 3. Efficient computation of the ground occupancy mask: the original view (on the left) is mapped to a plane through a combination of homographies that are chosen so that (1) verticality is preserved during projection from 3D scene to transformed view, and (2) ratio of heights between 3D scene and projected view is preserved for objects that lies on the same line in the transformed view.

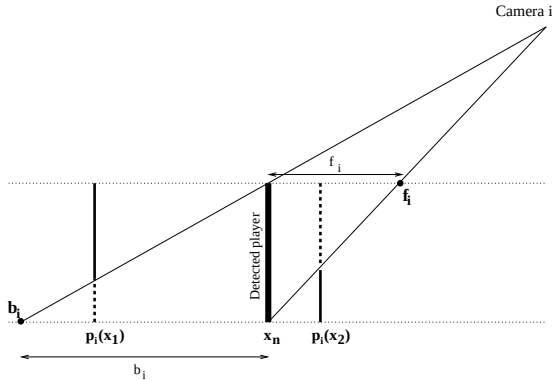


Fig. 4. Impact of occlusions on the update of ground occupancy mask associated to camera i . Dashed part of the vertical silhouette standing in $\mathbf{p}_i(\mathbf{x}_1)$ and $\mathbf{p}_i(\mathbf{x}_2)$ are known to be labeled as foreground since a player is known to be standing in $\mathbf{x}_n$. Hence they become useless to infer whether a player is located in $\mathbf{x}_1$ and $\mathbf{x}_2$, respectively.

ground position $\mathbf{x}$ that becomes non-informative in view i as a consequence of the presence of a player in $\mathbf{x}_n$. To estimate this ratio, we consider the geometry of the problem. Figure 4 depicts a plane $\mathcal{P}_i$ that is orthogonal to the ground, while passing through the i^{th} camera and the player position $\mathbf{x}_n$. In $\mathcal{P}_i$, we consider two points of interest, namely $\mathbf{b}_i$ and $\mathbf{f}_i$, which correspond to the points at which the rays, originated in the i^{th} camera and passing through the head and feet of the player, intersect the ground plane and the plane parallel to ground at height H , respectively. We denote f_i (b_i) to be the distance between $\mathbf{f}_i$ ($\mathbf{b}_i$) and the vertical line supporting player n in $\mathcal{P}_i$. We also consider $\mathbf{p}_i(\mathbf{x})$ to denote the orthogonal projection of $\mathbf{x}$ on $\mathcal{P}_i$, and let $d_i(\mathbf{x})$ measure the distance between $\mathbf{x}$ and $\mathcal{P}_i$. Based on those definitions, the ratio $\alpha_i(\mathbf{x}, \mathbf{x}_n)$ is estimated by

$$\alpha_i(\mathbf{x}, \mathbf{x}_n) = [(\delta - \min(\|\mathbf{p}_i(\mathbf{x}) - \mathbf{x}_n\|, \delta)) / \delta] \cdot [1 - \min(d_i(\mathbf{x}) / W, 1)] \quad (3)$$

with δ being equal to f_i or b_i , depending on whether $\mathbf{p}_i(\mathbf{x})$ lies ahead or behind $\mathbf{x}_n$, with respect to the camera. In (3), the first and second factors reflect the misalignment of $\mathbf{x}$ and $\mathbf{x}_n$ in $\mathcal{P}_i$ and orthogonally to $\mathcal{P}_i$, respectively.

Given $\alpha_i(\mathbf{x}, \mathbf{x}_n)$, the ground occupancy mask and aggregation weight of the i^{th} camera in position $\mathbf{x}$ are updated as follows:

$$G_i^{n+1}(\mathbf{x}) = \max(G_i^n(\mathbf{x}) - \alpha_i(\mathbf{x}, \mathbf{x}_n)G_i^0(\mathbf{x}_n), 0) \quad (4)$$

$$w_i^{n+1}(\mathbf{x}) = \max(w_i^n(\mathbf{x}) - \alpha_i(\mathbf{x}, \mathbf{x}_n), 0) \quad (5)$$

For improved computational efficiency, we limit the positions $\mathbf{x}$ investigated in the refined approach to the 30 local maxima that have been detected by the naive approach.

For completeness, we note that the above described update procedure omit the potential interference between occlusions caused by distinct players in the same view. However, the consequence of this approximation is far from being dramatic, since it ends up in, without affecting the information that is actually exploited. Taking those interferences into account would require to back-project the player silhouettes in each view, thereby tending towards a computationally and memory expensive top-down approach such as the one presented in [7].

Moreover, it is worth mentioning that, in a top-down context, the authors in [1] or in [7] propose formulations that simultaneously search for the K positions that best explain the multiple foreground masks observations. However, jointly considering all positions increases the dimensionality of the problem, and dramatically impact the computational load. Since our experimental results show that our proposed method does not suffer from the usual weaknesses of greedy algorithms, such as a tendency to get caught in bad local minima, we believe that

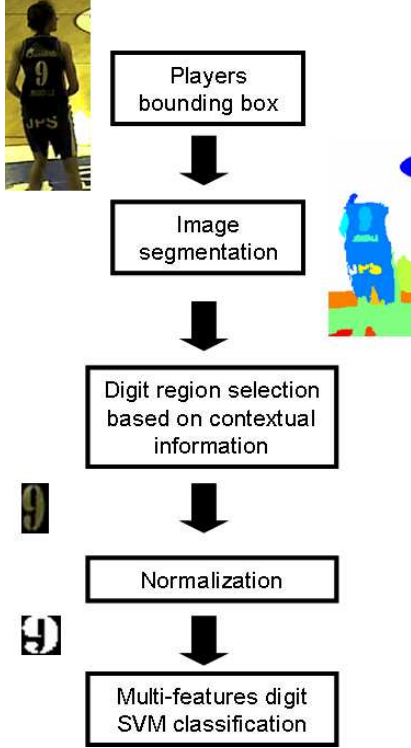


Fig. 5. Recognition of digits printed on players' shirts through segmentation, selection, and classification of regions that are likely to represent digits.

it compares very favorably to any joint formulation of the problem, typically solved based on iterative proximal optimization techniques.

IV. PLAYERS DIGIT RECOGNITION

This section considers the recognition of the digital characters printed on the sport shirts of athletes. The proposed approach is depicted in Figure 2. For each detected position on the ground plane, a $0.8\text{m} \times 2\text{m}$ conservative bounding box is projected in each one of the views. Each box is then processed according to an approach that is similar to the coarse-to-fine method introduced in [12]. In the initial step, the bounding box image is segmented into regions. Digit candidate regions are then filtered out based on contextual attributes. Eventually, selected regions are classified into '0-9' digits or bin classes, and the identity of the player is defined by majority vote, based on the results obtained in different views. Our proposed approach differs from [12] in the way each one of those steps is implemented.

Our segmentation step is based on the mean-shift algorithm [6], which is a pattern recognition technique that is particularly well suited to delineate denser regions

in some arbitrarily structured feature space. In the mean-shift image segmentation, the image is typically represented as a two-dimensional lattice of 3-dimensional $L \times u \times v$ pixels. The space of the lattice is known as the spatial domain, while the color information corresponds to the range domain. The location and range vectors are concatenated in a joint spatial-range domain, and a multivariate kernel is defined as the product of two radially symmetric kernels in each domain, which allows for the independent definition of the bandwidth parameters h_s and h_r for the spatial and range domains, respectively [6]. Local maxima of the joint domain density are then computed, and modes that are closer than h_s in the spatial domain and h_r in the range domain are pruned into significant modes. Each pixel is then associated with a significant mode of the joint domain density located in its neighborhood. Eventually, spatial regions that contain less than M pixels are eliminated. In our case, since there is a strong contrast between digit and shirt, we can afford a high value for h_r , which is set to 8 in our simulations. The parameter h_s trade-offs the run time of segmentation and subsequent filtering and classification stages. Indeed, a small h_r value defines a smaller kernel, which makes the segmentation faster but also results in a larger number of regions to process in subsequent stages. In our simulations, h_r has been set to 4, while M has been fixed to 20.

To filter out regions that obviously do not correspond to digits, we rely on the following observations:

- Valid digit regions never touch the border of the (conservative) bounding box;
- Valid digit regions are surrounded by a single homogeneously colored region. In practice, our algorithm selects the regions for which the neighbors of the 4 extreme (top/bottom, right/left) points of the region belong to the same region;
- The height and width of valid regions ranges between two values that are defined relatively to the bounding box size. Since the size of the bounding is defined according to real-world metrics, the size criterion implicitly adapts the range of height and width values to the perspective effect resulting from the distance between the detected object and the camera.

For completeness, it is worth mentioning that some particular fonts split some digits in two distinct regions. For this reason, candidate digit regions are composed of either a single or a pair of regions that fulfill the above criteria.

The (pairs of) regions that have been selected as eligible for subsequent processing are then normalized and

classified. Normalization implies horizontal alignment of the major principal axis, as derived through computation of moments of inertia, and conversion to a 24×24 binary mask. Classification is based on the 'one-against-one' multi-class SVM strategy [8], as recommended and implemented by the LIBSVM library [3]. A two-class SVM is trained for each pair of classes, and a majority vote strategy is exploited to infer the class (0 to 9 digit or bin class) from the set of binary classification decisions. In practice, to feed the classifier each region sample is described by a 30-dimensional feature vector, namely:

- 1 value to define the number of holes in the region;
- 3 values corresponding to second order moments m_{02} , m_{20} , and m_{22} ;
- 2 values to define the center of mass of the region;
- 2×12 values to define the histogram of the region along vertical and horizontal axis.

Numbers with two digits are reconstructed based on the detection of two adjacent digits. To run our simulations, we have trained the SVM classifier based on more than 200 manually segmented samples of each digit, and on 1200 samples of the bin class. The bin class samples correspond to non-digit regions that are automatically segmented in one of the views, and whose size is consistent with the one of a digit.

V. DETECTED PLAYERS TRACKING

To track detected players, we have implemented a rudimentary whilst effective algorithm. The tracks propagation is currently done over a 1-frame horizon, based on the Munkres general assignment algorithm[11]. Gating is used to prevent unlikely matches, and a high level analysis module is used to link together partial tracks using shirt color estimation. In the future, graph matching techniques should be used to evaluate longer horizon matching hypothesis. More sophisticated high level analysis should also be implemented, e.g. to exploit the available player recognition information or to duplicate the partial tracks that follow two players that are very close to each other.

VI. EXPERIMENTAL VALIDATION

A. Player detection and tracking

To evaluate our player detection algorithm, we have measured the average missed detection and false detection rates over 180 different and regularly spaced time instants in the interval from 18:47:00 to 18:50:00, which corresponds to a temporal segment for which a manual ground truth is available. This ground truth information consists in the positions of players and referees in the coordinate reference system of the court. We consider

that two objects cannot be matched if the measured distance on the ground is larger than 30 cm. Figure 6 presents several ROC curves, each curve being obtained by varying the detection threshold for a given detection method. Three methods are compared, and for each of them we assess our proposed algorithm to mitigate false detections. As a first and reference method, we consider the approach followed by [9], [10], which projects the foreground masks of all views only on the ground plane. The poor performance of this latter approach is mainly due to the shadows of the players, and to the small contribution of players' feet to the foreground masks. To validate this interpretation, in the second method, we have projected the foreground masks on a single plane located one meter above the ground plane. Doing this, the shadows influence is drastically attenuated, whilst the main contribution now originates from the body center parts, which are usually well represented in the foreground masks. We observe significant improvements compared to [9], [10]. The third and last detection method presented in Figure 6 is our proposed method. We observe that the benefit obtained from our ground occupancy integration is striking. The improvement brought by our false alarm detector is also quite obvious. In addition, the cross in Figure 6 presents an operating point achieved after rudimentary tracking of detected positions. We observe that taking into account temporal consistency can still further improve the detection results.

In the APIDIS setup, all areas of the basket court are not covered by the same number of cameras. Figure 7 shows the influence of the camera coverage on the missed and false detections rates. It also shows that in the areas with high coverage, most of the missed detections are due to players standing very close one to another.

B. Player recognition

To validate the player recognition pipeline, we have selected 190 bounding boxes of players from side views cameras. In each selected bounding box, the digit was visible and could be read by a human viewer, despite possibly significant appearance distortions. Table I summarizes our recognition results. The recognition rate is above 73%. More interestingly, we observe that when the digit was not recognized, it was most often assigned to the bin class, or did not pass the contextual analysis due to segmentation error. Moreover, the remaining 4% of false positive do not include any real mismatch between two digits. Actually, 75% of false positives were due to the miss of one digit in a two-digits number. In other cases, two digits have been recognized, the correct one and a false detected one.

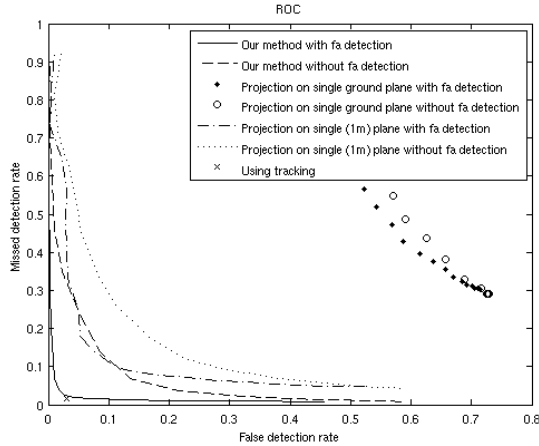


Fig. 6. ROC analysis of player detection performance.

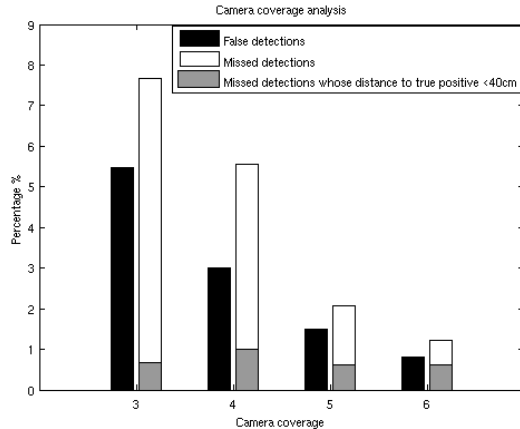


Fig. 7. Player detection performance wrt camera coverage.

Besides, a more detailed analysis has revealed that most of the non-recognized players were standing on the opposite side of the field, compared to the camera view from which the bounding box was extracted. In this case, the height of the digit decreases to less than 15 pixels, which explains the poor recognition performance, below 50%. In contrast, a camera located on the same side of the field than the player achieves close to 90% correct recognition rate.

Based on those observations, we are reasonably confident that the recognition performance of our system will be sufficiently good to assign a correct label to short segments of player trajectories, thereby providing a valuable tool both to raise tracking ambiguities or to favor a preferred player during video summary production.

Recognition	73 %
Segmentation error	11 %
False negative	12 %
False positive	4 %

TABLE I
PLAYER RECOGNITION PERFORMANCE.

VII. CONCLUSION

We have presented video processing algorithms to define the position and identity of athletes playing on a sport field, surrounded by a set of loosely synchronized cameras. Detection relies on the definition of a ground occupancy map, while player recognition builds on pre-filtering of segmented regions and on multi-class SVM classification. Experiments on the APIDIS real-life dataset demonstrate the relevance of the proposed approaches.

REFERENCES

- [1] A. Alahi, Y. Boursier, L. Jacques, and P. Vanderghenst, "A sparsity constrained inverse problem to locate people in a network of cameras," in *Proceedings of the 16th International Conference on Digital Signal Processing (DSP)*, Santorini, Greece, July 2006.
- [2] J. Berclaz, F. Fleuret, and P. Fua, "Principled detection-by-classification from multiple views," in *Proceedings of the International Conference on Computer Vision Theory and Application (VISAPP)*, vol. 2, Funchal, Madeira, Portugal, January 2008, pp. 375–382.
- [3] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," in <http://www.csie.ntu.edu.tw/~cjlin/papers/libsvm.pdf>.
- [4] F. Chen and C. De Vleeschouwer, "A resource allocation framework for summarizing team sport videos," in *IEEE International Conference on Image Processing*, Cairo, Egypt, November 2009.
- [5] —, "Autonomous production of basket-ball videos from multi-sensored data with personalized viewpoints," in *Proceedings of the 10th International Workshop on Image Analysis for Multimedia Interactive Services*, London, UK, May 2009.
- [6] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, May 2002.
- [7] F. Fleuret, J. Berclaz, R. Lengagne, and P. Fua, "Multi-camera people tracking with a probabilistic occupancy map," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 2, pp. 267–282, February 2008.
- [8] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, March 2002.
- [9] S. Khan and M. Shah, "A multiview approach to tracing people in crowded scenes using a planar homography constraint," in *Proceedings of the 9th European Conference on Computer Vision (ECCV)*, vol. 4, Graz, Austria, May 2006, pp. 133–146.
- [10] A. Lanza, L. Di Stefano, J. Berclaz, F. Fleuret, and P. Fua, "Robust multiview change detection," in *British Machine Vision Conference (BMVC)*, Warwick, UK, September 2007.
- [11] J. Munkres, "Algorithms for the assignment and transportation problems," in *SIAM J. Control*, vol. 5, 1957, pp. 32–38.
- [12] Q. Ye, Q. Huang, S. Jiang, Y. Liu, and W. Gao, "Jersey number detection in sports video for athlete identification," in *Proceedings of the SPIE, Visual Communications and Image Processing*, vol. 5960, Beijing, China, July 2005, pp. 1599–1606.