Tennis player segmentation for semantic behavior analysis

Vito Renò
CNR ISSIA
Via Amendola 122 D/O
70126 Bari, Italy
reno@ba.issia.cnr.it

Nicola Mosca
CNR ISSIA
Via Amendola, 122 D/O
70126 Bari, Italy
mosca@ba.issia.cnr.it

Massimiliano Nitti
CNR ISSIA
Via Amendola, 122 D/O
70126 Bari, Italy
nitti@ba.issia.cnr.it

Tiziana D'Orazio
CNR ISSIA
Via Amendola, 122 D/O
70126 Bari, Italy
dorazio@ba.issia.cnr.it

Donato Campagnoli
Mas-Tech srl
Via Cantone, 96
41032 Cavezzo (MO), Italy
campagnolidonato@gmail.com

Andrea Prati
Università IUAV di Venezia
D.P.P.A.C - Santa Croce 1957
30135 Venezia, Italy
aprati@iuav.it

Ettore Stella
CNR ISSIA
Via Amendola, 122 D/O
70126 Bari, Italy
stella@ba.issia.cnr.it

Abstract

Tennis player silhouette extraction is a preliminary step fundamental for any behavior analysis processing. Automatic systems for the evaluation of player tactics, in terms of position in the court, postures during the game and types of strokes, are highly desired for coaches and training purposes. These systems require accurate segmentation of players in order to apply posture analysis and high level semantic analysis. Background subtraction algorithms have been largely used in sportive context when fixed cameras are used. In this paper an innovative background subtraction algorithm is presented, which has been adapted to the tennis context and allows high precision in player segmentation both for the completeness of the extracted silhouettes. The algorithm is able to achieve interactive frame rates with up to 30 *frames per second, and is suitable for smart cameras embedding. Real experiments demonstrate that the proposed approach is suitable in tennis contexts.*

1. Introduction

In recent years, there has been an increasing interest by the scientific community to develop technology platforms which provide coaches with solutions that allow them to more effectively train the next generation of athletes [11]. Automatic and objective analysis of athletes performance during competitions is very important for coaches and training strategies. In particular in the tennis context, the positions of players during the game, the reactions to different events and the sequences of strokes are very important to evaluate performance and match results. In all these situations, visual systems for the automatic player behavior analysis require an accurate segmentation of the silhouette in order to allow more complex semantic processing.

In the tennis context some commercial systems are available for real time ball detection [3] and manual annotation of video sequences [1, 2, 5] either off-line or in real-time. A system for player segmentation in broadcast images has been published in [10] which uses useful information of the context like the uniform court color and white court lines for semantic analysis such as instant speed and speed change of the player, as well as the positions of the player in the court. In [8] a platform for extracting semantic information from multi-camera tennis data is presented. The player tracking process is evaluated only in terms of blob positions that are compared with a UbiSense tag-tracking system. The system presented in [15] archives action summaries in the form of 3D motion trajectories of the essential elements of the game (players, ball). The tennis player tracking system presented in [13] uses an improved CamShift algorithm which is initialized by an interframe difference method. The main con-

straint of the approach is the use of the strong difference of colors between players and the court to extract the player. In [6] motion and color information are used to extract areas of activity by accumulating motion information over all frame pixels and over several frames, and by color information of the background to select pixels belonging to moving objects. The resulting information can be used to extract semantic data such as the kind of shot in a tennis context.

A common consideration in many algorithms is that the color of the court is uniform, as well as the surrounding area [21]. These features allow to separately build background models for the field inside the court and the surrounding area, instead of creating a complete background for the whole image [14]. In order to overcome the problems related to poor segmentation, wearable sensors such as lightweight inertial sensors, have been chosen for stroke recognition and skill assessment. Detecting spikes in the accelerometer data provides information on the impact of the ball on the tennis racket and the temporal location of tennis strokes [9]. Anyway this kind of sensors are not allowed during official matches, and also athletes are quite hostile to wear invasive sensors.

In this paper we approached the challenge of tennis player segmentation with vision cameras by considering two main aspects: on one hand the development of a robust algorithm for the detection of the player silhouette in its most complete shape as possible to allow more complex semantic analysis; on the other hand privileging the design of an efficient algorithm in terms of computational load for supporting real time applications.

In the tennis context some specific issues limit the application of standard approaches and require ad hoc solutions. First of all, the segmentation of players in clay-court or synthetic court: the color of the player skin or player uniform can be confused with the color of the court and the segmentation fails in the detection of the whole silhouette. Moreover, indoor tennis courts are generally characterized by lighting conditions that combine the complexities found in both outdoor and indoor contexts: during daylight, the illumination depends on sunny or cloudy conditions (the tensile structures are actually semi transparent); when artificial light is on, the light flickering affects camera acquisitions. Last but not least, the usage of high frame rate cameras, necessary for reliable ball tracking, discourages the usage of complex algorithms for the extraction of players silhouettes, if real time performance need to be achieved.

In this paper we propose a background subtraction approach to address these specific challenges with respect to the tennis context. Starting from the analysis and the results obtained by standard and well assessed approaches such as the Adaptive Mixture of Gaussians (MoG) [20] [23], non-parametric models as the GMG [12], adaptive light-weight algorithms [7] [22] and the adaptive background estimation and foreground detection using kalman-filtering [18], an adaptive BG model able to deal with high frame rate videos and dynamic scenes has been developed. The analysis of the variance of each grey level has been done to model the sensor response to different light intensities. A blob analysis has been applied to both the difference between the background model and the current frame, and the temporal difference between consecutive frames. These steps are useful to extract robust foreground areas (moving players and ball) while maintaining a low computational load. Experimental results demonstrate the effectiveness of the proposed approach when compared with standard methods.

The rest of the work is organized as follows: in the second section the proposed algorithm is detailed, the third section contains experiments and results and the last one discusses the conclusions and future works.

2. Methodology

2.1. System Overview

The proposed algorithm, internally referred as GIVE-BACK (Globally Intrinsic VariancE for BACKground), is the preliminary step of a system aimed to address coaching needs. For this reason, it is designed to operate in indoor environments, although it can operate outdoor as well. The proposed architecture consists of four cameras that are placed at the corners of the court to cover all the game areas with at least two views. Cameras position and corresponding images are shown in Figures 1 and 2. It is worth noting that overlapping views are exploited to produce a synchronized broadcast video while no information are passed through different cameras. Therefore, the proposed approach is not dependant on multiple views (gains

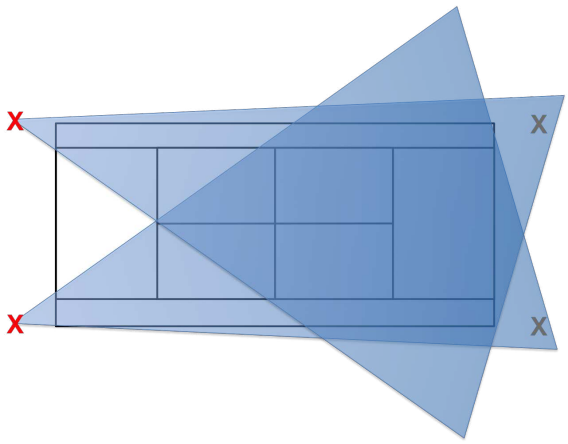

Figure 1. The figure depicts the position of the acquisition hardware on the tennis court, highlighted with red marks. Cameras (X) are put at the corners of the external rectangle, so that each pair can acquire the opposite part of the court.

Figure 2. Example of synchronized acquisition. Four frames are captured exactly at the same time to make the system capable of observing the whole court from at least two different points of view.

scalability) and can run both on single or multi camera systems.

2.2. Algorithm Description

GIVEBACK pipeline can be divided in three main building blocks, as shown in Listing 1: initialization, processing and update. The first step is executed only once and initializes the BG image setting each pixel to half intensity. Therefore, the produced image is gray and reflects the absence of any *a priori* knowledge about the scene. The processing phase is composed of: variance, one step frame differencing, fine tuning and energy. The ones marked with an asterisk in Listing 1 are the same presented in [16], while the others are detailed singularly in the following sub sections. The BG image is updated according to PIIB logic [16] enriched by a binary update mask M_{upd}. Hence, each BG pixel value is increased or decreased by κ if the corresponding M_{upd} value is set to true (in our implementation $\kappa = 1$). Finally, the fine tuning phase exploits the output of two blob analyses — one on the foreground mask and the other on the one step frame differencing one — with the aim of giving robustness to the BG model, as it will be described later.

Listing 1. Algorithm pseudocode

```
Background Initialization*
for each frame
    Variance process
    One step frame differencing
    if(Background is learned)
        Foreground extraction
        Fine tuning process
    Background Update*
    Energy Process*
```

2.3. Variance Process

The variance model exploited in this work is useful to describe different sensor responses to different light intensities, so that its value is not related to the observations of a single pixel over time, but is a function of a specific gray level in a byte range [17]. Therefore, for each frame, the location of the occurrences of each generic gray value γ is first stored in a set

$$\text{Obs}(\gamma) = \{k = (u, v) | BG_{t-1}(u, v) = \gamma\} \tag{1}$$

Then, the variance V at the time t, associated to the γ-th gray level is iteratively updated with the following formula:

$$V_t(\gamma) = \frac{V_{t-1}(\gamma) \cdot N_{t-1}(\gamma) + \sum_k |I_t(k) - BG_{t-1}(k)|^2}{N_t(\gamma)} \tag{2}$$

where $k \in \text{Obs}(\gamma)$, $N(\gamma)$ is the number of times the γ-th gray level occurred over time and BG is the background. In the equations BG is substituted with the latest available frame (I_{t-1}) while the BG is being learned, namely until the energy gradient descent reaches its minimum value.

2.4. One step frame differencing

This task is executed at each iteration and produces a binary mask obtained by thresholding the absolute difference of the last captured frame and the one being processed. First, the absolute difference image is calculated with the formula:

$$AD = |I_t - I_{t-1}| \tag{3}$$

Then, for each pixel (u, v) the binary mask M_{os} is calculated in the following way:

$$M_{os} = \begin{cases} 0 & \text{if } AD(u, v) \leq \tau\left(I_{t-1}(u, v)\right) \\ 255 & \text{if } AD(u, v) > \tau\left(I_{t-1}(u, v)\right) \end{cases} \tag{4}$$

Each pixel is considered as a normal random variable: the mean value is represented by its corresponding value in the last captured frame, while the variance depends on its gray level, since different intensity values might have different variances. The threshold $\tau(\cdot)$ used to classify each pixel as background or foreground is a function of a specific gray value and in our implementation it is set to $\tau(\gamma) = 3.5\sigma_\gamma$, where $\sigma_\gamma = \sqrt{V(\gamma)}$. Hence, each backgound pixel (in black) lies in an interval $[\gamma - 3.5\sigma_\gamma, \gamma + 3.5\sigma_\gamma]$ while the foreground ones (in white) represent the tails of the corresponding normal distribution. The binary mask obtained at this stage is useful to achieve robustness during the subsequent phases, for example avoiding the BG model update in correspondence of a moving player.

2.5. Foreground extraction

The foreground extraction phase is similar to the one step frame differencing one, except from the fact that the background image is exploited instead of the last captured frame.

The output of this module is a binary mask M_{fg} obtained with the same thresholding process presented in Eq. 4, in which the absolute difference image is $AD = |I_t - BG_{t-1}|$. M_{fg} is the mask used to compare the model with other approaches in the next section.

2.6. Fine Tuning Process

After the BG model has been learned by the system, the fine tuning process module is switched on to calculate the binary update mask M_{upd}. This task is achieved by means of a blob analysis done on both M_{os} and M_{fg} to obtain two sets of connected regions — namely $B_{os} = \{b_1, b_2, \ldots, b_n\}$ and $B_{fg} = \{b_1, b_2, \ldots, b_m\}$ — that are processed according to the following rule:

$$M_{upd} = \{\ulcorner(b_i, b_j)\urcorner | b_i \in B_{os}, b_j \in B_{fg}, b_i \cap b_j \neq \emptyset\} \quad (5)$$

where $\ulcorner(b_i, b_j)\urcorner$ is the minimum circumscribed rectangle that embeds both b_i and b_j. Each region extracted from the foreground mask is compared to each region extracted by the one step frame differencing process in order to find overlapping blobs that do not produce an empty set when intersected. As a consequence, the update mask keeps trace of robust foreground areas in which the BG update does not take place, allowing the algorithm to easily filter ghosts or static subjects that stand still on the scene.

3. Experiments and Results

Two variants of the methodology described in the previous section, under the name of GIVEBACK and GIVEBACK fine tuned, have been tested and compared with other statistical based background models available in the BGS library [19] (GMG and MOGv2) and the adaptive background estimator based on kalman filtering [18] implemented in MVTec Halcon suite [4]. The first variant of the proposed algorithm models the background skipping the fine tuning process, while the complete method — with the fine tuning process in place — is tested separately as well.

Both qualitative and quantitative tests have been done on recorded sequences that represent a tennis training session. Four raw videos have been taken with AVT Prosilica GT1920C cameras capable of acquiring 1936×1456 frames at $40Hz$ and configured to capture 1920×1024 frames at $50Hz$ in order to avoid flickering issues exploiting the hardware setup. Moreover, cameras are equipped with auto iris lenses which enable to ensure a constant brightness level in the whole recordings. As a consequence, results obtained on a single camera are reproducible on the other ones when recording the same event from different points of view as represented in Figure 4.

Table 1 shows the performance of each step of GIVEBACK. This implementation is capable of running at 30 fps, since variance process can be stopped after about 256 iterations. This is valid when considering the computationally expensive process that is used in the fine tuned version of GIVEBACK, while better frame rates can be achieved by the plain version of the proposed algorithm. These results have been obtained on an Intel Xeon E5-2603 @ 1.60 GHz, 32GB RAM, Windows 7 64bit OS.

Table 1. GIVEBACK performance evaluated for each step of the algorithm

Task	**Elapsed time**	
	[ms]	[%]
Variance process	12	29
Foreground extraction	1	3
Fine tuning process	27	64
Background update	1	2
Energy process	1	3

Experiments have been conducted in the following way: starting from a reference frame f_0, ten images sampled every 500 frames have been manually annotated and then quantitatively analysed exploiting the corresponding ground truth masks. Only moving players and balls have been segmented on the ground truth image, while inactive balls (always present in tennis courts, especially during training sessions) have not been annotated as foreground objects.

Qualitative results in terms of player silhouette segmentation can be inferred from the visual inspection of the foreground objects resulting from different algorithms, as reported in Figure 3. Here, GMG algorithm handles effortlessly shadows near the players feet and shows a tendency to consider background some parts of the legs, performing poorly on the lower parts of the player body because of color similarity between the court and the player skin. Kalman fitering based backgound estimator is sensitive to ghosting issues that appear when the player moves after having stationed elsewhere. The proposed approach is able to produce a well-cut player silhouette, especially in the fine tuned variant where the ghost is being reduced while preserving the whole shape of the player.

Figure 5 summarizes the algorithms performance in terms of Precision P and Recall R for each annotated frame. Here, each point in the $P - R$ plane refers to a run of a specific background subtraction method where different algorithms are shown with different marker shapes and colors, while variants are presented as color-filled or white-filled. According to this representation the ground truth has coordinates $(1, 1)$, therefore points that lie in the upper right part of the figure correspond to the best results.

A brief overview of the chosen metrics is given here: let TP be the number of true positives pixels, FP be the number of false positives pixels, TN be the number of true nega-

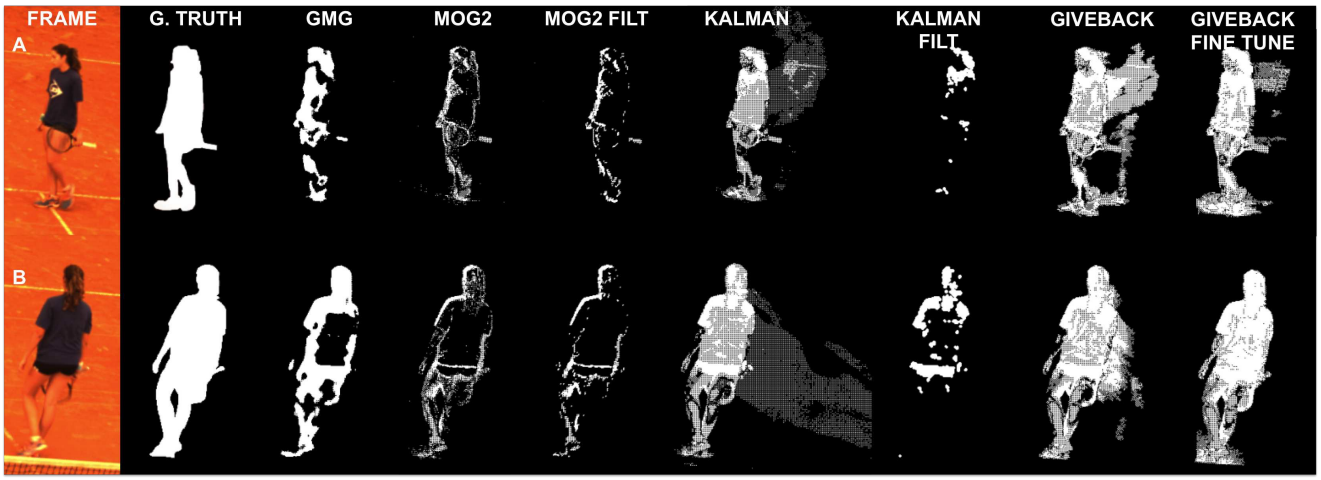

Figure 3. Example of silhouettes processed with different background algorithms. A player is shown from two different point of views. Each row, from left to right, shows the starting image, the manually annotated mask, and the masks obtained with the GMG, MOGv2 and Kalman based background algorithms. The results from the two variants of the GIVEBACK approach are shown in the last columns. Where salt and pepper noise is visible, a "filtered" variant is tested as well. The GIVEBACK fine tuned is the algorithm that better preserves the entire player silhouette with a low computational load. The amount of false positive or negative pixels in the proposed approach is reduced when compared to the other statistical methods considered.

tives pixels and FN be the number of false negatives pixels on the foreground mask. Accordingly, P, R and F-Measure F are defined as:

$$P = \frac{TP}{TP+FP}, R = \frac{TP}{TP+FN}, F = 2 \cdot \frac{P \cdot R}{P+R} \quad (6)$$

In the comparison, both MOGv2 and the Kalman filter based background FG masks have been post processed with a morphological opening operation employing a circular structuring element of 2 pixels radius. The GMG algorithm did not require any additional filtering operation since the method already produces salt-and-pepper noise filtered foreground masks.

Figure 5 shows that both the variants proposed in this paper have noticeable performance. The variant related to the fine tuned algorithm shows the best overall results, with higher average scores on both axes (better precision and recall performance at the same time). MOGv2 is particularly sensitive to salt-and-pepper noise and has a global tendency to show low recall values. This implies a high number of false negatives pixels in the FG masks, as it can be seen analysing the silhouettes in Figure 3. The Kalman filtering based background results (highlighted by blue diamonds) in terms of precision are not constant during the acquisition. This means that the approach is affected by the production of false positive pixels in the form of player ghosts, as shown in Figure 3. Finally, GMG algorithm shows a precision comparable with the reference performance of the Kalman based one, trading some precision for better recall scores.

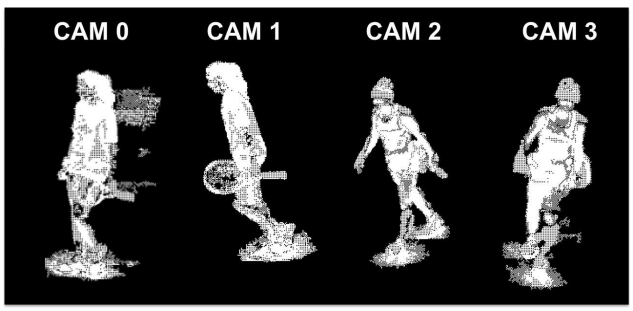

Figure 4. Example of player silhouettes extracted from four syncronized views. CAM0 and CAM1 refer to Player 1, while CAM2 and CAM3 to Player 2. The performance of the proposed approach is similar across all views, showing the robustness of the method.

In some respects, the GMG algorithm and the "complete" variant of the GIVEBACK algorithm described in this paper perform similarly well, with the GMG algorithm being better in the precision score and the ones proposed here showing better recall. However, as will be shown shortly, the adaptive BG model presented here seems more reliable, with a uniform behavior while working on different frames, while scores obtained by the GMG algorithm are more scattered.

Figure 6 shows a boxplot of the F-Measure calculated during the experiment. There are seven boxes, one for each algorithm. Inside each box, the median value is highlighted with a red line, while the edges of the box are the 25^{th} and 75^{th} percentiles. The whiskers extend to the most extreme data points not considered outliers, and outliers are marked

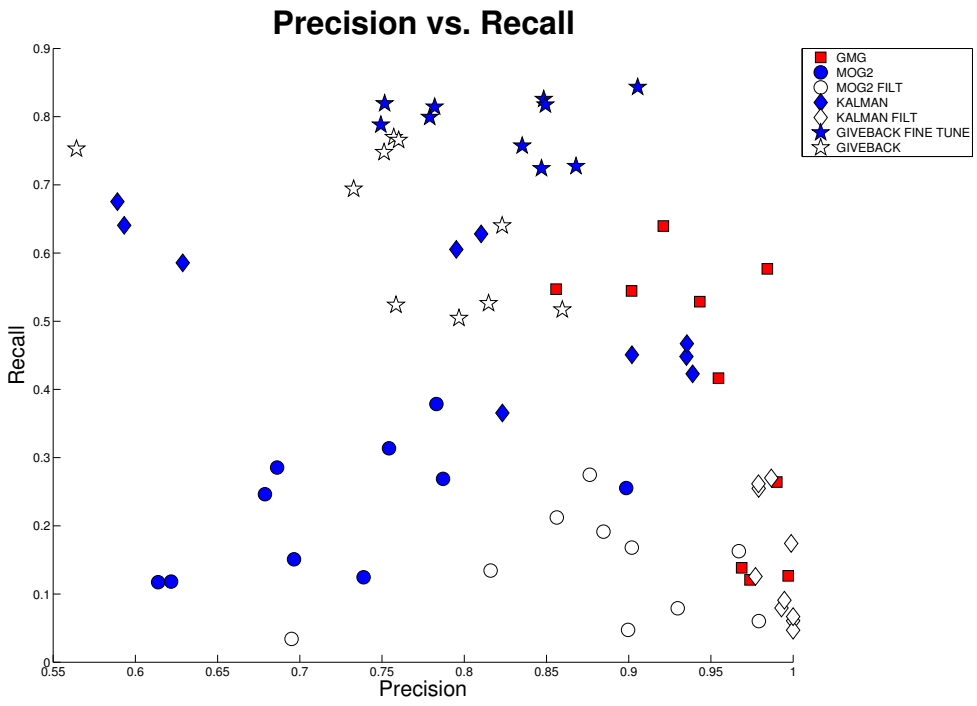

Figure 5. Quantitative results on the dataset in terms of Precision and Recall. Each point corresponds to a comparison between a foreground mask obtained with a specific algorithm and the corresponding ground truth. The upper right corner represents a FG mask that that is exactly the same as the ground truth (both P and R values equal 100%). Points that tend to $(1, 1)$ are the best among the considered ones.

individually with a red cross. Hence, small boxes refer to algorithms whose results are repeatable over time, while big boxes show that the range of F is wide (reflecting high variance in the results). The only algorithm that produces outliers is the Kalman filter-based one, due to not constant precision values among the executions as highlighted beforehand. However, it is the one with the smaller box.

In summary, the best algorithm among the ones tested is the proposed method enriched by the fine tuning module, as its median F value is 80%, the associated box is the second smallest and there are no outliers in the statistic.

4. Conclusion and future works

In this paper, an efficient method to segment active entities — players and balls — in tennis context is presented. The proposed approach is based on simple but effective operations (from a computational load point of view) that allow its employment on real time systems. Moreover, it operates directly on raw videos thus encouraging its implementation directly on smart cameras. Experiments on tennis training video sequences demonstrate its effectiveness in tennis players silhouettes processing, even if usually there is a strong similarity between players skin and the tennis court. The fine tuned version of the algorithm shows good scores in terms of Precision and Recall and F-Measure. Its performance on different frames are very similar on each ground truth annotated test image. These results confirm the robustness of the proposed method when compared to other statistical approaches evaluated in the benchmark. Moreover, foreground masks extracted by the GIVEBACK algorithm better preserves players silhouettes thus enabling high level analysis such as posture recognition. Future works will be directed forward to semantic analysis based on player silhouettes, to embedding the algorithm on smart devices along with further optimizations.

References

[1] Avenir sports. *http://avenirsports.ie/*.

[2] Dartfish. *http://www.dartfish.com/en/index.htm*.

[3] Hawk-eye innovations official website. *http://www.hawkeyeinnovations.co.uk/*.

[4] Mvtec halcon. *http://www.halcon.com/*.

[5] Performa sports. *http://www.performasports.com/*.

[6] A. Briassouli, V. Mezaris, and I. Kompatsiaris. Color aided motion-segmentation and object tracking for video sequences semantic analysis. In *International Journal of Imaging Systems and Technology - Special Issue on Applied Color Image Processing*, volume 17, pages 174–189, 2007.

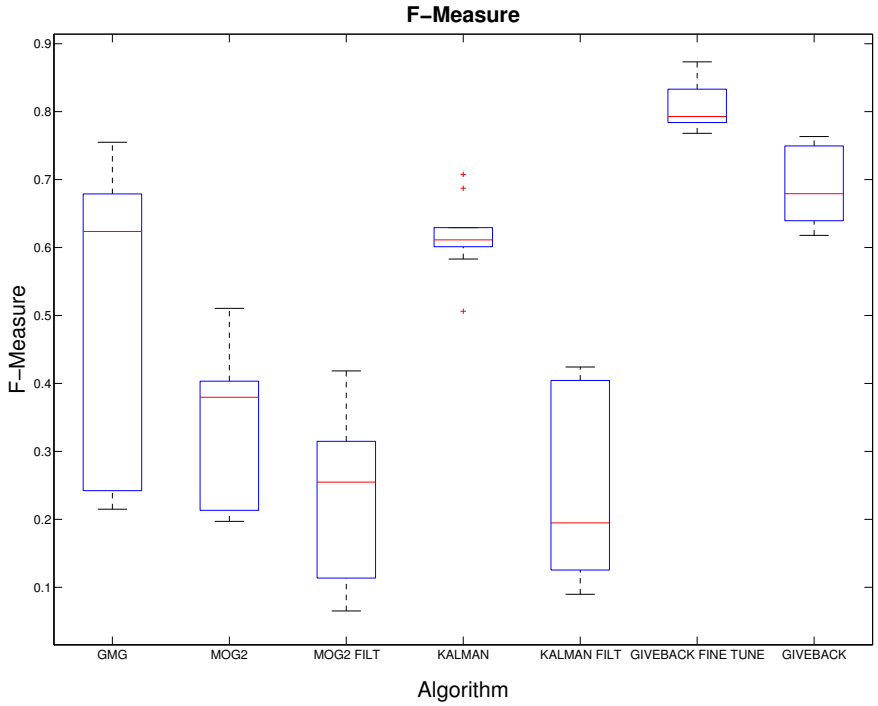

Figure 6. Quantitative results in terms of F-Measure organized in a boxplot. The figure summarizes the overall performance of each algorithm. The red marks represent the median value among the executions, while blue boxes go from the 25^{th} to the 75^{th} percentile. Small boxes are better because of low variance and high repeatability of the experiments.

[7] M. Casares, S. Velipasalar, and A. Pinto. Light-weight salient foreground detection for embedded smart cameras. *Computer Vision and Image Understanding*, 114(11):1223 – 1237, 2010. Special issue on Embedded Vision.

[8] C. Ó. Conaire, P. Kelly, D. Connaghan, and N. E. O'Connor. Tennissense: A platform for extracting semantic information from multi-camera tennis data. In *Digital Signal Processing, 2009 16th International Conference on*, pages 1–6. IEEE, 2009.

[9] D. Connaghan, P. Kelly, N. O'Connor, M. Gaffney, M. Walsh, and C. O'Mathuna. Multi-sensor classification of tennis strokes. In *IEEE sensors*, pages 1437–1440, 2011.

[10] B. Dang, A. Tran, T. Dinh, and D. Thang. A real time player tracking system for broadcat tennis video. In *Intelligent Information and Database Systems*, pages 105–113. Springer,LNAI 5991, 2010.

[11] T. D'Orazio and M. Leo. A review of vision-based systems for soccer video analysis. *Pattern recognition*, 43(8):2911–2926, 2010.

[12] A. Godbehere, A. Matsukawa, and K. Goldberg. Visual tracking of human visitors under variable-lighting conditions for a responsive audio art installation. In *American Control Conference (ACC), 2012*, pages 4305–4312, June 2012.

[13] Y. Guo, S. Lao, and L. Bai. Player detection algorithm based on color segmentation and improved camshift algorithm. In *Proceedings of the 2012 International Conference on Information Technology and Software Engineering*, 2013.

[14] J. Han, D. Farin, and P. de With. Broadcast court-net sports video analysis using fast 3-d camera modeling. *IEEE Trans. on Cirtuis abd System for Video Technology*, 18(11):1628–1638, 2008.

[15] G. Pingali, Y. Jean, and I. Carlbom. Lucent vision: A system for enhanced sports viewing, volume 1614 of. *Lecture Notes in Computer Science*, pages 689–696.

[16] V. Renò, R. Marani, T. D'Orazio, E. Stella, and M. Nitti. An adaptive parallel background model for high-throughput video applications and smart cameras embedding. In *Proceedings of the International Conference on Distributed Smart Cameras*, ICDSC '14, pages 30:1–30:6, New York, NY, USA, 2014. ACM.

[17] V. Renó, R. Marani, N. Mosca, M. Nitti, T. D'Orazio, and E. Stella. A likelihood-based background model for real time processing of color filter array videos. In *New Trends in Image Analysis and Processing–ICIAP 2015 Workshops*, pages 218–225. Springer, 2015.

[18] C. Ridder, O. Munkelt, and H. Kirchner. Adaptive background estimation and foreground detection using kalman-filtering. In *Proceedings of International Conference on recent Advances in Mechatronics*, pages 193–199. Citeseer, 1995.

[19] A. Sobral. BGSLibrary: An opencv c++ background subtraction library. In *IX Workshop de Visão Computacional (WVC'2013)*, Rio de Janeiro, Brazil, Jun 2013.

[20] C. Stauffer and W. Grimson. Learning patterns of activity using real-time tracking. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):747–757, Aug 2000.

[21] K. Teachabarikiti, T. Chalidabhongse, and A. Thammano. Players tracking and ball detection for an automatic tennis video annotation. In *11th Int. Conf. Control, Automation, Robotics and Vision*, pages 2491–2494, 2010.

[22] X. Yu and D. Farin. Current and emerging topics in sports video processing. In *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on*, pages 526–529, July 2005.

[23] Z. Zivkovic. Improved adaptive gaussian mixture model for background subtraction. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 2, pages 28–31 Vol.2, Aug 2004.