# COMPUTATIONAL UAV CINEMATOGRAPHY FOR INTELLIGENT SHOOTING BASED ON SEMANTIC VISUAL ANALYSIS

*Fotini Patrona, Ioannis Mademlis, Anastasios Tefas, and Ioannis Pitas*

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
{fotinip, imademlis, tefas, pitas}@aiia.csd.auth.gr

## ABSTRACT

Audiovisual coverage of sports events using Unmanned Aerial Vehicles (UAVs) is becoming increasingly popular. Intelligent audiovisual (A/V) shooting tools, accurately identifying the 2D region of cinematographic attention (RoCA) depicting rapidly moving target ensembles and automatically controlling the UAVs/cameras through visual content analysis, are thus needed. A novel algorithmic pipeline is proposed, implementing computational UAV cinematography for assisting sports coverage, based on semantic, human-centered visual analysis. Athlete and ball detection / tracking results as well as their spatial distribution on the image plane are the semantic features extracted from UAV video feed and exploited for RoCA extraction, based solely on present and past target detections. A PID controller visually controlling a real or virtual camera to track the RoCA and produce aesthetically pleasing shots, without exploiting 3D location-related information, is employed. The proposed method is evaluated on actual UAV footage from soccer matches and promising results are obtained.

***Index Terms—*** autonomous UAVs, cinematography, sports broadcasting, human-centered visual analysis, PID controller

## 1. INTRODUCTION

As employing camera-equipped Unmanned Aerial Vehicles (UAVs) for audiovisual coverage of sports events tends to become mainstream, new cinematography challenges arise that need to be properly handled [1, 2]. Such a case is the automatic region-of-cinematographic-attention (RoCA) identification for guiding autonomous UAV camera framing, mimicking the way a human camera operator would do. Most research works implementing camera control based on computational aesthetics study low-level video features, like texture, saturation and hue [3, 4], while overlooking the influence of higher-level visual semantics (e.g., athlete location and framing on the image plane), which can be acquired through visual scene (stadium) and targets (athletes) analysis, providing information concerning the visual attention of a professional cameraman when shooting. Especially for sports shooting, such a high-level visual analysis is crucial, since there usually are multiple moving targets, e.g., athletes. Only by considering them to be parts of one unified RoCA and by modeling its global structural motion, can the UAV camera be controlled to best capture the complex game dynamics.

An autonomous quadcopter capable of capturing frontal images of moving targets by selecting the best vantage points is presented in [5]. A Partially Observable Markov Decision Process (POMDP) is employed for estimating target motion intentions and deciding between moving or staying still, thus minimizing camera motion, while it always faces the target. High-level cinematographic commands concerning the shots to be captured and the desired object positioning are provided in [6], where UAV navigation is performed by implementing smooth transitions between the requested shots, while simultaneously tracking the targets.

Fixed cameras are employed in [7] for tracking basketball players along with the ball and estimating player centroid. Path planning and control are then performed by utilizing a virtual camera, thus proposing a hybrid, low-latency system incorporating knowledge about future events. Future regions-of-interest are estimated in [8], based on the stochastic field representing the motion trends of soccer players. The robustness of this spatiotemporal video content selection method is also evaluated with moving cameras, and their motion control is demonstrated to mimic the one of a cameraman.

Aesthetics criteria, e.g., player spatial distribution, ball visibility and game flow are taken into consideration in the automated director assistance method proposed in [9]. The camera capturing the most visually appealing shots is identified and personalized directorial shooting styles are learnt, while various visual semantics, e.g., intensity of activity, object detection and tracking, object size, width, height, orientation and location, as well as motion vectors, are some of the features employed by the best-view selection system [10].

The methods described above utilize future information in order to perform camera control (e.g., by buffering video

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No731667 (MULTIDRONE). This publication reflects the authors' views only. The European Commission is not responsible for any use that may be made of the information it contains.

frames) and focus on capturing specific targets/objects. On the contrary, the proposed pipeline implements automatic camera control and can be employed both during production and post-production stages, as it relies only on present and past visual information, in order to estimate a RoCA.

In cases of target ensembles, like sports games, objective definition of the RoCA and proper framing are rather difficult. Tracking the ball and focusing on its trajectory in order to define the RoCA over the game field view, would be one solution. Human-centered visual analysis, i.e., player detection/tracking and 2D spatial player distribution, for calculating the region enclosing the majority of them, while omitting outliers, would be another, in case ball detections are not available.

The proposed pipeline employs the combination of ball-/player-related information and aesthetics criteria for estimating the parameters determining the shots to be produced. Moreover, instead of robotic cameras, employed by the vast majority of the related approaches, the proposed pipeline introduces the use of autonomous UAVs for sports coverage, drastically reducing (post-)production cost, since very low-cost UAVs can be easily found and no cameramen or UAV operators are required - the vehicle may simply hover over the field sideline, with its camera set to the widest possible field of view.

In brief, the proposed pipeline consists of three stages, and the only input required is a UAV video frame. At first, player and ball detection/tracking are performed, with the resulting regions-of-interest (ROIs) subsequently used by the RoCA estimation component. The latter calculates the video frame region of the most interest, employing the rule-of-thirds [11] while also taking ROI motion direction into account, and feeds a special PID controller designed for cinematographic camera control [12] with the estimated RoCA. Finally, the PID controller guids a real camera gimbal or a virtual camera appropriately, aiming to keep the RoCA focused and centered. Virtual camera control is simply and efficiently implemented by suitably cropping the appropriate region of the original video frame. An informative and aesthetically pleasing (real or virtual) output video frame is thus produced.

The main novelties introduced by the proposed pipeline could be summarized as follows: a) framing based on a RoCA and not a specific object/target (e.g., game player, ball), b) exploitation of present and past information only in order to form the RoCA trajectory (no knowledge of the future required), and c) camera control based solely on 2D visual information - no usage of 3D location-related information concerning the camera or the target.

## 2. A COMPUTATIONAL UAV CINEMATOGRAPHY PIPELINE

The proposed pipeline operates on a UAV camera video stream, by extracting high-level player-related information and controlling a virtual pan-tilt-zoom camera, so that the most interesting RoCA of the entire field view is always properly framed and visualized. In order to operate appropriately, the method requires high-resolution, wide-angle long shots of the field. Modern 2K and 4K video cameras are thus considered ideal, since they have high enough resolution to allow the proposed algorithm pipeline to produce high-quality virtual camera video frames, being spatially cropped segments of the original video frames. The execution pipeline presented in Fig. 1, is thoroughly described in the following paragraphs.

To begin with, player and ball detection/tracking have to be performed, using any player/ball detector (e.g., [13]) and 2D visual tracker (e.g., [14]), which normally output image ROIs defined by their top-left and bottom-right corners, expressed in pixel coordinates. The simplest solution for target detection and localization is to use an existing pedestrian/ball detector, e.g., [15] based on CNNs, possibly finetuned with case-specific image data sets. Semantic information of this sort can nowadays be relied upon, thanks to advances in deep learning.

Let us denote by $f_t$ the UAV video frame being processed at time instance $t$, its ball ROI by $\mathcal{R}_{bt} = [x_{bmin}, y_{bmin}, x_{bmax}, y_{bmax}]^T$ and its player ROIs by $\mathcal{R}_{t,i} = [x_{min}, y_{min}, x_{max}, y_{max}]^T$, $i = 1, \ldots, N$, with $N$ denoting the players involved in the game. The mean distance $\overline{dx}_{t,i}$ of player $i$ from his $n = 3$ nearest neighbors, along the $x$ axis, is calculated. Afterwards, having sorted player ROIs in ascending linear order along the $x$ direction, the Euclidean distance of $\mathcal{R}_{t,i}$ from $\mathcal{R}_{t,i\pm1}$ regarded as 1D points, is calculated, and two checks are performed: a) if it is greater than the mean player ROI size in this direction (i.e., mean width in $x$ direction) and b) if it is greater than the mean distance $\overline{dx}_{t,i}$. In case the two previous tests succeed, the player is considered an outlier [16], thus being omitted, and the previous steps are repeated for the next player, until the first one to be included in the RoCA is found. The entire procedure is subsequently performed along the $y$ axis. This way, a bounding box $RoCA_t$ enclosing the RoCA at time instance $t$ is calculated, containing the players distributed in the most serried way. However, if the obtained RoCA is smaller in size than a user-specified percentage of the original video frame, its size is adjusted accordingly. This way, extreme close-up shots are avoided.

If ball ROI $\mathcal{R}_{bt}$ information is available, it is represented by its middle point $cR_{bt} = [c_{bx}, c_{by}]^T$ and its motion direction along axes $x, y$ is estimated. Representing player ROIs by their center point, as well, the ROI centroid $cR_t = [c_x, c_y]^T$ is calculated along with its direction of motion on both axes $x, y$. After smoothing $cR_t$, by applying a Gaussian filter of temporal length $L_{cR} = 3$ video frames on its coordinates, $RoCA_t$ is estimated in such a way that camera fixation point $fp_t = f_{Rb} * cR_{bt} + (1 - f_{Rb}) * cR_t$, $f_{Rb} = 0.75$ always constitutes one of the four intersection points arising by the cinematographic rule-of-thirds. In case ball detection

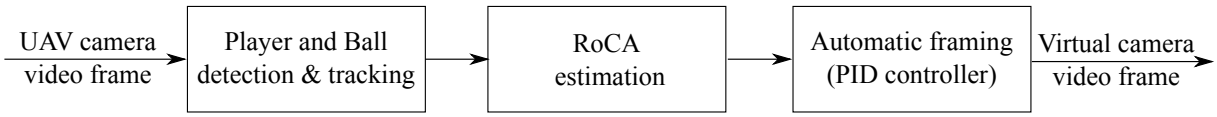



**Fig. 1**. Proposed method pipeline

is missing at some time instance $t$, the fixation point is calculated based on the last detected ball ROI, and $f_{Rb}$ weight is assigned to $cR_t$.

More specifically, by exploiting motion direction information and $RoCA_t$ size (previously calculated), its bounding box is formed around $fp_t$ so that $\frac{2}{3}$ of the RoCA size always lie towards the ball motion direction (or the players' if ball detection is not available) and only $\frac{1}{3}$ towards the opposite direction. $RoCA_t$ coordinates are also filtered by a Gaussian window of temporal length $L_{RoCA} = 5$ video frames, so that a pleasing trajectory, without abrupt movements and jumps, can be obtained. Finally, its aspect ratio is estimated, and the smallest modifications required for retaining $16 : 9$ ratio are performed. It is then fed to the PID controller [12] that is responsible for controlling the pan, tilt and zoom of the virtual camera appropriately, so that $RoCA_t$: a) always encloses at least a user-specified percentage of the original video frame, b) always covers a user-specified percentage of the produced virtual camera video frame, i.e., a specific cinematographic shot type is captured, c) is appropriately positioned in the virtual camera video frame, based on user preferences.

The PID controller proposed in [12] is employed here as an image-based virtual camera control system implementing a number of the previously identified cinematographic shot types [17, 18, 19, 20], requiring information concerning neither the UAV, nor the target 3D position. Several target-tracking camera rotation types can be defined as a set of requirements relating 2D visual information and camera orientation. By exploiting these requirements, this purely vision-based controller can instantly control virtual camera pan and tilt, thus effectively executing a target-tracking shot based solely on 2D visual information. The controller keeps the target RoCA properly positioned within the resulting virtual video frame and framed according to the desired shot type [11], by appropriately modifying the virtual camera zoom. Either central composition or the rule of thirds can be followed for RoCA framing [1]. Unless otherwise stated, central composition is always used in this work.

Fig. 2 presents indicative results obtained through this pipeline. To elaborate, Fig. 2(a) depicts the original UAV video frames with the ball bounding box painted in yellow, the player bounding boxes in red, the enclosing RoCA in green, and the virtual video frame window estimated by the PID controller in blue, while the resulting virtual video frame is shown in Fig. 2(b). It can be easily noticed that the players positioned further apart were considered outliers and were not included in the estimated RoCA.

## 3. EXPERIMENTAL EVALUATION

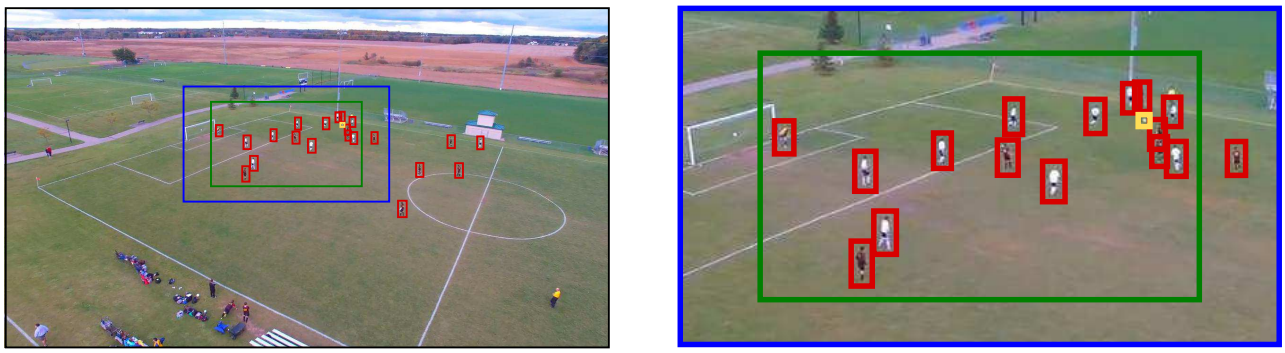
Performance assessment of the proposed pipeline was performed during post-production, employing non-professional UAV footage of a soccer game as input data, captured at a resolution of $1280 \times 720$ pixels. A video clip of 4335 video frames was selected for presenting evaluation results. Ground truth ROIs for the players and the ball were manually annotated, while a professional cameraman was asked to indicate a RoCA ground truth sequence. Player ROIs were also automatically extracted, with the aid of YOLO v3 detector [15].

The proposed pipeline, as well as method [7] and especially virtual camera handling, were both implemented in Python. Comparisons of the obtained evaluation results with ground truth and automatically detected player ROIs are provided. Moreover, the RoCAs estimated by the two methods are compared to the human expert-defined sequences.

The following user-defined parameters were employed. Central composition and a virtual video frame coverage percentage of 70%, with a minimum RoCA size equal to $\frac{1}{3}$ of the original video frame, i.e., $427 \times 240$ pixels were the parameters concerning RoCA formation and PID control. The Gaussian window sizes adopted were $L_{cR} = 3, L_{RoCA} = 5$, both selected by visually inspecting the resulting virtual videos. The virtual camera fixation point $fp_t$ was positioned in between the ball ROI middle point $cR_{bt}$ and player centroid $cR_t$, with $cR_{bt}$ weight equal to $f_{Rb} = 0.75$ and $cR_t$ weight $f_{Rt} = 1 - f_{Rb} = 0.25$ on axis $x$, so that the virtual camera can follow horizontal ball movement. In the contrary, $f_{Rb} = 0.25$ and $f_{Rt} = 0.75$ on the $y$ axis, thus minimizing the influence of vertical ball position to virtual camera movement and preventing the latter from following ball bounce.

Fig. 3 presents the percentage of the original UAV video frame region characterized as RoCA by the human expert, the proposed pipeline and method [7], employing both ground truth (gt) player ROIs and YOLO detections. Contrary to the [7]-created RoCA, the RoCA produced by the proposed pipeline does not appear to differ significantly from the ground truth. This, can be also verified by Table 1, presenting mean coverage percentage of the original video frame by the RoCAs produced in all the aforementioned cases along with standard deviation values, as the mean coverage achieved by the proposed pipeline is much closer to the ground truth RoCA coverage. In addition, the reported standard deviation values highlight the robustness of the proposed method, as they are dramatically lower than the ones estimated for [7].

With a mean RoCA coverage of the original video frame



(a) (b)

**Fig. 2**. (a) UAV original camera-captured video frame, (b) virtual camera video frame – yellow box: ball, red boxes: players, green box: estimated RoCA, blue box: PID controller framing window

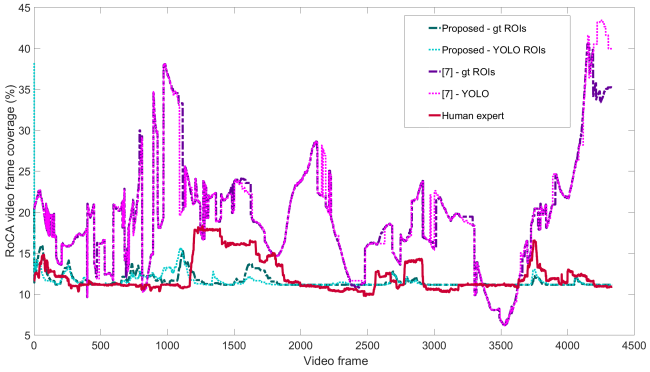



**Fig. 3**. Original video frame coverage (%) by RoCA – dashed dark green line: proposed with ground truth player ROIs, dotted cyan line: proposed with YOLO detected ROIs, dashed deep purple line: [7] with ground truth player ROIs, dotted magenta line: [7] with YOLO detected ROIs, solid red line: human expert ground truth

equal to 12.16%, according to the human expert-created ground truth RoCA sequence, the vast majority of the video frames having RoCAs extending only in 10 - 18% of their image and no single video frame having a RoCA covering more

**Table 1**. Mean RoCA coverage of the original video frames and standard deviation results

| | mean coverage & std (%) |
|---|---|
| Proposed - gt ROIs | **11.57±0.80** |
| Proposed - YOLO ROIs | 11.51±0.79 |
| [7] - gt ROIs | 19.65±6.73 |
| [7] - YOLO ROIs | 19.80±7.24 |
| Human expert | **12.16±2.00** |

than 20% of the entire image, it becomes obvious that quite a lot of inactive visual 2D space was originally captured, which is reasonable since the UAV was capturing long-shots.

**Table 2**. Average Intersection over Union results

| | mean IoU |
|---|---|
| Proposed - gt ROIs | 54.30% |
| Proposed - YOLO ROIs | **54.39%** |
| [7] - gt ROIs | 44.21% |
| [7] - YOLO ROIs | 43.23% |

Another metric employed for the evaluation of the obtained RoCAs, is their average instersection over union (IoU) with the ground truth. As shown in Table 2, the proposed pipeline outperforms [7] in terms of IoU by approximately 10%. Additionally, the results obtained with ground truth and YOLO-detected player ROIs for both methods do not differ significantly, which is due to YOLO effectiveness.

## 4. CONCLUSIONS

In this paper, a pipeline for computational cinematography in UAV sports coverage, applicable both during production and post-production, based on semantic, human-centered visual analysis was proposed. High-level semantic features, like player/ball detection/tracking results (ROIs), as well as player spatial distribution and motion direction, are extracted from an aerial video feed of a single UAV and, after performing visual analysis, the RoCA is extracted using the rule-of-thirds, based solely on present and past information. The RoCA is then operated upon by a suitable PID controller, which can visually control a real or virtual camera to track the target and produce salient video shots, keeping it properly framed by appropriately cropping the original video frames, without using 3D location-related information. Promising results were obtained by the objective evaluation of the proposed pipeline, which outperformed a competing method.

## 5. REFERENCES


[1] I. Mademlis, V. Mygdalis, N. Nikolaidis, and I. Pitas, "Challenges in autonomous UAV cinematography: An overview," in *Proceedings of IEEE International Conference on Multimedia and Expo (ICME)*, Jul 2018, pp. 1–6.

[2] I. Mademlis, N. Nikolaidis, A. Tefas, I. Pitas, T. Wagner, and A. Messina, "Autonomous unmanned aerial vehicles filming in dynamic unstructured outdoor environments," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 147–153, 2018.

[3] M. Nishiyama, T. Okabe, Y. Sato, and I. Sato, "Sensation-based photo cropping," in *Proceedings of the 17th ACM International Conference on Multimedia*. 2009, MM '09, pp. 669–672, ACM.

[4] P. Obrador, L. Schmidt-Hackenberg, and N. Oliver, "The role of image composition in image aesthetics," in *2010 IEEE International Conference on Image Processing*, Sept 2010, pp. 3185–3188.

[5] R. Coaguila, G.R. Sukthankar, and R. Sukthankar, "Selecting vantage points for an autonomous quadcopter videographer.," in *FLAIRS Conference*, 2016, pp. 386–391.

[6] Q. Galvane, J. Fleureau, F.L. Tariolle, and P. Guillotel, "Automated cinematography with unmanned aerial vehicles," *CoRR*, vol. abs/1712.04353, 2017.

[7] P. Carr, M. Mistry, and I. Matthews, "Hybrid robotic/virtual pan-tilt-zom cameras for autonomous event recording," in *Proceedings of the 21st ACM International Conference on Multimedia*. 2013, MM '13, pp. 193–202, ACM.

[8] K. Kim, D. Lee, and I. Essa, "Detecting regions of interest in dynamic scenes with camera motions," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1258–1265.

[9] C. Chen, O. Wang, S. Heinzle, P. Carr, A. Smolic, and M. Gross, "Computational sports broadcasting: Automated director assistance for live sports," in *2013 IEEE International Conference on Multimedia and Expo (ICME)*, July 2013, pp. 1–6.

[10] F. Daniyal, M. Taj, and A. Cavallaro, "Content and task-based view selection from multiple video streams," *Multimedia Tools and Applications*, vol. 46, no. 2, pp. 235–25, Jan 2010.

[11] I. Tsingalis, A. Tefas, N. Nikolaidis, and I. Pitas, "Shot type characterization in 2D and 3D video content," in *2014 IEEE 16th International Workshop on Multimedia Signal Processing (MMSP)*, Sept 2014, pp. 1–5.

[12] N. Passalis, A. Tefas, and I. Pitas, "Efficient camera control using 2D visual information for unmanned aerial vehicle-based cinematography," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, May 2018, pp. 1–5.

[13] D. Triantafyllidou, P. Nousi, and A. Tefas, "Fast deep convolutional face detection in the wild exploiting hard sample mining," *Big Data Research*, vol. 11, pp. 65–76, 2018.

[14] H.K. Galoogahi, A. Fagg, and S. Lucey, "Learning background-aware correlation filters for visual tracking," *CoRR*, vol. abs/1703.04590, 2017.

[15] J. Redmon, S. Kumar Divvala, R.B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *CoRR*, vol. abs/1506.02640, 2015.

[16] I. Pitas and A.N. Venetsanopoulos, *Nonlinear Digital Filters*, Kluwer international series in engineering and computer science: VLSI, computer architecture, and digital signal processing. Springer US, 1990.

[17] I. Karakostas, I. Mademlis, N. Nikolaidis, and I. Pitas, "UAV cinematography constraints imposed by visual target tracking," in *Proceedings of IEEE International Conference on Image Processing (ICIP)*, Oct 2018, pp. 76–80.

[18] I. Mademlis, V. Mygdalis, C. Raptopoulou, N. Nikolaidis, N. Heise, T. Koch, J. Grunfeld, T. Wagner, Messina. A., F. Negro, et al., "Overview of drone cinematography for sports filming," in *European Conference on Visual Media Production (CVMP)*, 2017.

[19] I. Mademlis, V. Mygdalis, N. Nikolaidis, M. Montagnuolo, F. Negro, A. Messina, and I. Pitas, "High-level multiple-UAV cinematography tools for covering outdoor events," *IEEE Transactions on Broadcasting (accepted for publication)*, 2019.

[20] O. Zachariadis, V. Mygdalis, I. Mademlis, N. Nikolaidis, and I. Pitas, "2D visual tracking for sports UAV cinematography applications," in *Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Nov 2017, pp. 36–40.