Building a labeled dataset for recognition of handball actions using mask R-CNN and STIPS

Ivašić-Kos, Marina; Pobar, Miran

Source / Izvornik: **2018 7th European Workshop on Visual Information Processing (EUVIP), 2019, 1 - 6**

Conference paper / Rad u zborniku

Publication status / Verzija rada: **Accepted version / Završna verzija rukopisa prihvaćena za objavljivanje (postprint)**

https://doi.org/10.1109/EUVIP.2018.8611642

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:195:898413

Rights / Prava: In copyright / Zaštićeno autorskim pravom.

Download date / Datum preuzimanja: **2024-07-23**

Sveučilište u Rijeci
Fakultet informatike i digitalnih tehnologija

Repository / Repozitorij:

Repository of the University of Rijeka, Faculty of Informatics and Digital Technologies - INFORI Repository

DIGITALNA KNJIŽNICA

DIGITALNI AKADEMSKI ARHIVI I REPOZITORIJI

Building a labelled dataset for recognition of handball actions using Mask R-CNN and STIPS

Marina Ivašić-Kos
Department of Informatics
University of Rijeka
Rijeka, Croatia
marinai@uniri.hr

Miran Pobar
Department of Informatics
University of Rijeka
Rijeka, Croatia
mpobar@uniri.hr

***Abstract*— Building successful machine learning models depends on large amounts of training data that often needs to be labeled manually. We propose a method to efficiently build an action recognition dataset in the handball domain, focusing on minimizing the manual labor required to label the individual players performing the chosen actions. The method uses existing deep learning object recognition methods for player detection, and combines the obtained location information with a player activity measure based on spatio-temporal interest points to track players that are performing the currently relevant action, here called active players. The method was successfully used on a challenging dataset of real-world handball practice videos, where the leading active player was correctly tracked and labelled in 84% of cases.**

Keywords— object detectors; sports scenes; Mask R-CNN, spatio-temporal interest point – STIP; action recognition database

I. INTRODUCTION

Object detection and action recognition are problems of supervised machine learning for which it is crucial that knowledge is automatically acquired through experience. Simplified, the prerequisite for supervised learning is that data is in the form (x, y) where the input value is an example x and the target value is a class y. A prerequisite for development of automatic action detection and recognition models is a video dataset with labeled action clips. The task of supervised learning in this case is to find a model or function f(x) = y, which assigns the action class y to an example video clip x. A large number of examples will contribute to the discovery of rules implicitly contained in the data, better generalization, and system performance improvements. But it is not easy to prepare a learning database for this type of task.

Deep learning based solutions have recently had great success in domains such as image classification and object detection, and one of the key factors contributing to the progress made in these fields is the availability of very large datasets for learning, for example the object detection dataset MS COCO contains over two million labeled objects in about 200 000 images [1].

The progress in these domains motivates the development of similar technologies for automatic video analysis, which is more complex as it combines the audio and spatial with temporal image components. One of the fundamental goals of video analysis is detection and recognition of actions performed by people in video, where detection entails temporal and spatial segmentation of the person performing the action, while recognition entails labelling the performed action (e.g. jumping, running, waving, etc.).

The labelling for action recognition is normally done manually and involves many working hours, which makes the development of learning data sets for action recognition an expensive, extremely laborious and time-consuming task. The annotators should view each video, determine the beginning and end of an action, cut clips and mark objects on frames. If the video contains multiple persons in the frame, each person performing the action should be marked in all frame as well. For comparison, over 22 working hours were needed per 1,000 segmentations of object instances during annotation /preparation of COCO image database [1].

For the early action recognition datasets such as KTH [2] and Weizmann [3] the temporal segmentation could be done manually in a reasonable amount of time since the scenes were recorded with a static camera in controlled conditions with one person performing an action over a homogeneous background. In that case, the researchers could employ background subtraction or Chroma keying to segment the single actor from the background and then process the mask.

Nowadays, the datasets are much larger and more complex, videos are recorded under uncontrolled conditions, with more persons simultaneously on the scene performing various actions. A simple solution for segmentation and mask detection such as background subtraction was no longer a good solution, so labelling the video clips and marking objects with bounding boxes become predominantly manual and even more dependent on human labor. Therefore, the most common approach to labelling a video dataset became the use of crowdsourcing services such as Amazon Mechanical Turk, in order to employ a large number of human annotators. To facilitate distributed labeling of actions in video, specialized online annotation tools were developed, e.g., [4,5]. LabelMe video [4] enables annotators to mark the relevant actor shapes only in certain key frames, while it uses interpolation in conjunction with global motion estimation to insert the missing annotations between key frames. The annotators can then review and manually edit the interpolated shapes across different frames. An iterative approach is used in [6], where a classifier is trained on a small number of labeled frames to assist in annotating the remaining data, which is then in turn used to improve the classifiers. Still, even with crowdsourcing the problem of annotating video data

This research was fully supported by Croatian Science Foundation under the project IP-2016-06-8345 “Automatic recognition of actions and activities in multimedia content from the sports domain” (RAASS).

XXX-X-XXXX-XXXX-X/XX/$XX.00 ©20XX IEEE

remains challenging, as it was found [5] that the majority of crowdsourced workers produce unreliable work that has to be extensively validated, and the cost is still significant.

A different approach is taken in [7], where a game engine is used to generate a synthetic dataset of action recognition videos as an alternative to labeling large amounts of real-world data. This dataset is then used to augment a real-world dataset. An advantage of this approach is that large amounts of relatively varied data can be generated procedurally without a requirement for manual segmentation of actors.

In this paper, we present an approach that strives to minimize the manual labor in labelling an action recognition learning dataset, in the domain of sports videos. The idea is to leverage the existing object recognition methods to detect the players in the court, and then to determine active players, i.e. the ones performing the currently relevant action. For these reasons we have proposed the MR-CNN+STIPs method that is based on Mask R-CNN [8] for person detection and the space-time interest points [9, 10] (STIP) that jointly combine space and time information detected within the person bounding boxes for selection of active players. This can greatly simplify the creation of the action dataset, as only the start and end times of each action should be manually labeled, while the most time-consuming part of marking the active player is done automatically.

The rest of the paper is organized as follows: in Section II. The handball sports domain is explained. In Section III. The proposed method that combines Mask R-CNN video object detector and spatio-temporal interest points to determine the active players in handball scenes, is described. In Section IV. we have examined its performance on a custom dataset consisting of indoor and outdoor handball scenes recorded during handball school. The paper ends with a conclusion and the proposal for future research.

II. HANDBALL DOMAIN

Handball is a dynamic team sport. In team sports like handball, multiple players are present in the scene at the same time. Handball players quickly change their positions, combine different actions and change roles in the game within offense or defense but although they all might move and interact, not all players contribute to the currently most relevant action.

Handball has well-defined rules of the game which includes a number of players on the field, allowed actions and techniques, but during training and lessons while developing and practicing handball techniques, coach or teacher usually modifies these rules to maintain a high activity level with fast technique change and more repetitions. For this reason, most activities are performed in parallel to keep the waiting time between the activities as short as possible and repeating until all players change and practice the technique. For example, when practicing the throwing techniques, there is one student who shoots the ball at the goal, goalkeeper that moves to save the goal, while others wait in the queue or collect their balls around the court after performing the activity and run to their position in the queue.

III. PROPOSED MR-CNN+STIP METHOD FOR BUILDING A LABELLED DATASET

The goal of the proposed method is to automatically determine active players in the scene to make the tracking and labeling of the actions easier. Active players are those that perform certain technique such as shooting, dribbling, passing the ball or practice some handball skills. They should be distinguished from inactive players that are also present on the scene but do not perform techniques of interest for learning the action model such as waiting in a queue, sitting on a bench, collecting the balls or being the audience.

An overview of the proposed method is shown in Fig. 1. For each frame, players on the scene are detected in parallel with points with significant variations in velocity and non-constant appearance between sequential frames that can correspond to the moving objects (STIPs). Players are marked with bounding boxes. Players who are moving more will have more movement and shape changes and therefore their bounding box will have a higher STIPs density and will be marked as active. The detected bounding boxes of active players are tracked across the whole sequence to form player trajectories.

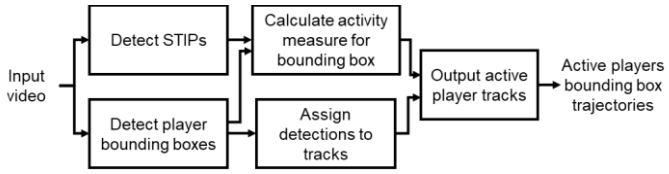

Fig. 1. An overview of the active player track detection.

A. Detection of players

To detect the players in the scene a convolutional neural network that is able to detect and localize objects, Mask R-CNN was used. Mask R-CNN [8] is an extension of Faster R-CNN [11] that follows a two-stage design in which candidate object bounding boxes or regions of interest (RoI) are proposed first and then in a sliding window fashion a deep fully convolutional network is applied to measure membership to object classes vs. background. The network generates object masks and bounding boxes for all possible classes with the corresponding confidence values and adds a parallel fully connected network branch for predicting segmentation masks on each RoI to make the selection of boxes and masks.

We have used Mask R-CNN in the standard Resnet-101-FPN network configuration with the parameters pre-trained on the COCO dataset, with no additional training on our dataset. In this experiment, only the bounding boxes and confidence values for the objects of the class "person" were used.

To eliminate false positive detections, we have considered only the bounding boxes with confidence scores higher than a given threshold. The threshold was experimentally determined to provide a good balance of high detection rate with few false positives and was set to 0.55. The output of the Mask R-CNN detector is shown in Fig. 2.

Fig. 2. Bounding boxes with confidence values as results of Mask R-CNN person detection

B. *Player tracking*

Since the Mask R-CNN operates on a given video frame independently of the others, it cannot automatically conclude which bounding boxes relate to the same objects in successive frames. Thus, an additional post-processing step is needed to track the players across the video and to obtain their trajectories.

In the first video frame, a track Id is assigned to each detected bounding box with confidence value higher than the threshold. Then, for each next frame, the assignment of the given detected bounding boxes to tracks is done using the Munkres' version of the Hungarian algorithm [12] which minimizes the total assignment cost of every detection to the track. The assignment cost of a given detection to every track depends on the location of the bounding box with regard to the track and the size difference between the box and the last known box in the track.

More formally, a new bounding box is determined to correspond to the track with minimal cost computed as a weighted sum of the Euclidean distance between the detected bounding box centroids (C_b) and the predicted track centroids ($C_{b+1}{}'$) considering the absolute area difference of the detected boxes area (P_b) and the last box area assigned to the track (P_{b-1}) (1):

$$(C_b, P_b) = \underset{b}{\operatorname{argmin}}\, w \sum_{b \in B} (d_2(C_b, C_{b+1}{}') + |P_b - P_{b-1}|); w \in [0,1], B \in N \quad (1)$$

The location of the bounding box in the previous frame is used as a prediction for the location in the next frame. Even though the players can move fast in the field and can switch the positions, this has performed rather well since the full frame rate of the source video is used. The use of Kalman filter to predict box location was also considered but in preliminary testing, but it has shown uneven performance and was not used further in the experiment.

In Fig. 3 examples are given of person bounding boxes with ID 20 through different frames and with corresponding confidence value obtained on these frames.

Since the detection of players is not perfect and players may enter or exit the camera field of view at any time, the number of tracks can change throughout the video, and some tracks should resume after a period where no detection has been assigned.

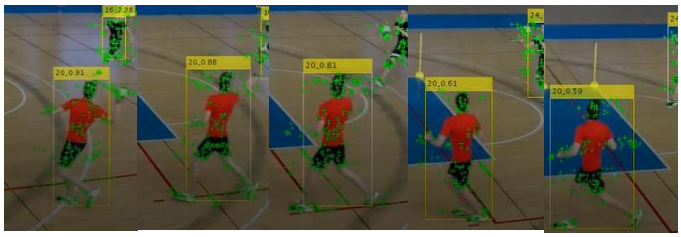

Fig. 3. In each video frame, a bounding box has its track ID

A cost threshold T is used to set the maximum allowed distance between a detected bounding box and track. A box whose cost of assignment to a track is greater than this threshold cannot be assigned to that track even though it might be the closest one to the track. If no detections are assigned to a track for M successive frames, no new detections are added to the track. The values of M and T are experimentally set to 20 and 100.

C. *3.3 Determining the active players*

The bounding boxes capture the location of the players but don't carry information about the movements of the players.

It is expected that actions in sports videos will be characterized by strong variations in velocity and appearance over time. To capture that information, spatio-temporal interest point detection is used. In the spatial domain, points with significant local variation of image intensities have been extensively investigated in the past and successfully applied in a number of applications such as optic flow estimation and tracking [13], image indexing and recognition [14]. The idea of spatial interest points is extended into the spatio-temporal domain by requiring the image values in space-time to have large variations in both the spatial and the temporal dimensions. Points with significant variations of image values in a local spatio-temporal neighborhood are denoted as “spatio-temporal interest points”, STIPs.

However, STIPs are local features that capture the spatio-temporal „interestingness" at different points in image regardless of whether those points correspond to the moving background or any object on the scene.

Therefore, the idea is to merge the information about player locations with local STIPs information in order to detect active players that are performing actions related to handball techniques among others that have already performed the desired action and are preparing for the next action by collecting balls in the court, running to next position in the game, waiting in the queue or sitting on the bench.

To combine the two pieces of information, for each detected bounding box (B) in each frame, an activity measure (A_b) per frame is computed using the density of spatio-temporal interest points (STIPs) in the same box (2):

$$A_b = \#stips \in \mathrm{B}/P_b, \quad (2)$$

where P_b is the area of the box B.

The STIPs are extracted using the selective STIPs method [10] with default parameters from the whole video. An example of STIPs detected on the whole image and the player bounding box are presented in Fig. 4.

Fig. 4. Player bounding box (white) and spatio-temporal interest points (green).

D. *3.4 Active player trajectories*

The detected bounding boxes are tracked across the whole sequence to form player trajectories, Fig.5. Active player score is then calculated as the average activity measure of the player along the trajectory. The result is a set of player trajectories with corresponding player activity scores.

Fig. 5. Tracking of active players across successive frames

An activity threshold can then be used to filter active from inactive players. As an alternative, when it is known that one player is active at a time in the video, i.e. only one player is performing an action, the player with max active player score can be selected as the leading player, Fig. 6.

Fig. 6. Detected leading player (white box) and her trajectory through the whole sequence (yellow line)

The result of the proposed method is a bounding box collage that contains all phases of an action and corresponds to the trajectory of the active player along with the action label. In Fig. 7. an example of active player collage for jump-shot action is shown. The collage contains a sequence of player bounding boxes presenting all phases of jump-shot action from running, take-off, flight, throw and landing.

Fig. 7. Active player collage for jump-shoot action.

IV. IMPLEMENTATION AND DISCUSSION OF LABELED DATASET BUILDING METHOD

We have tested the proposed method on a custom indoor and outdoor handball dataset consisting of footage acquired during practice and competition in a handball school. Handball games were organized for students on the whole or on a part of the court of the sports hall, but also on outdoor terrains so the background is cluttered, with challenging illumination, with a variable number of players and with other not ideal conditions. Hard shadows and reflections were often present as was some motion blur that made the problem even harder.

The dataset consisted of 600 videos in 4 action classes: dribbling, passing, shooting and jump-shoot. Multiple players appear in each video, about 10 in average, and each of them can perform an action. Recorded videos are cut in such a way to contain just one handball technique. Each file is manually labeled with a single "main" action of interest performed by one or more players at the same time, even though other players may perform actions such as running. In case of shooting, jump-shooting and dribbling the labeled action is most often performed by one leading player, but in the case of passing the ball two players are involved. The total duration of labeled actions is 1690s. The videos were shot indoors and outdoors using stationary GoPro cameras from different angles and in varying lighting conditions. In the indoor scenes, the cameras were mounted at a height of around 3.5 m to the left or right side of the field. For the outdoor scenes, the camera was at a height of 1.5 m. The videos were recorded with 30 frames per second in full HD resolution (1920x1080). Object detection with Mask R-CNN and STIPs extraction were both performed on full resolution videos and with no frame skip.

First, just the used Mask R-CNN object detector was tested on the dataset in isolation [15]. The detectors performance was evaluated in terms of recall, precision, and F1 score [16], counting the true positive detection when the intersection over union of the detected bounding box and the ground truth box exceeded the threshold of 0.5. The detector efficiency depends heavily on the number and size of objects on the scene, as well as the occlusion of objects, but are good enough to be used to build a learning set, Fig. 8. It could be noted that shadows and reflections of many objects resemble the actual object, but they were very rarely mistaken for the real object.

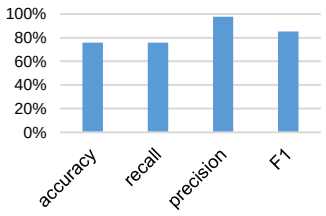

Fig. 8. Mask R-CNN results of player detection

Next, the active player detection was tested. We have considered for evaluation in terms of true positives only the cases when there is only one leading player in sequences. Examples of leading player detection are understood as true positives when the active leading player is correctly detected throughout the whole sequence (fig. 9, left). The true positive rate is the number of true positives divided by the number of tested sequences. The achieved average true positive rate for four handball action classes is 84%. In the majority of remaining cases, the players were tracked correctly but the wrong player was selected as active. In this case, it was only necessary to correct the active player id manually, while in the remaining cases the player was not tracked correctly and the example had to be either manually labeled or excluded from the database.

Fig. 9. Marking of the leading player (white thick bounding box); left is a correct detection and right is a wrong detection

For creating a learning set for supervised machine learning, it is important to collect as many well-labelled examples as possible. Our goal was to automatically build a dataset of labelled handball actions so that it can be used to train the action classifier. Therefore, it is important not only to generate as many examples of actions as possible, but to have consistent and well-defined examples. Actions are stored in a learning set as action collages that includes phases of a particular action and action name.

Examples of shooting action collages taken from a different view angle are shown in Figure 10. Each player should perform each of the technical elements according to game rules, but due to its morphological characteristics and motor abilities, the degree of specialization for the particular positions as well as the specific situations in the field each player has a specific performance of the same action.

In Figure 10 it is seen that the actions in the collage are shown with different number of thumbnails so that in some cases the same phase of action is shown in several frames. The number of thumbnails depend on the performance of player detector at all phases of the action performance and player activity during the tracking.

Due to the game being dynamic and a large number of active players, there are realistic situations on game footage where the players overlap so the built dataset includes also collages with occlusions, Fig. 11.

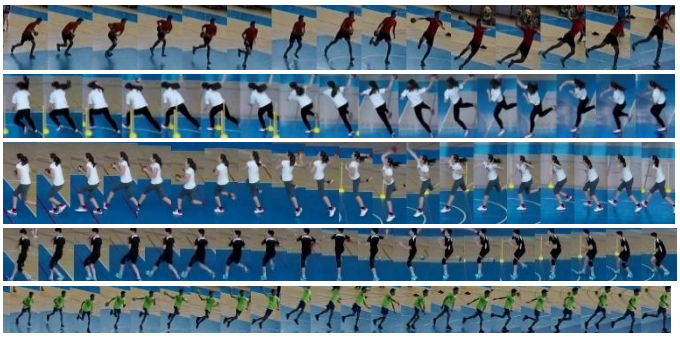

Fig. 10. Examples of shooting action collages

Fig. 11. Examples of occlusion in action collages

In addition to actions of interest, actions of players who do not perform the default action are also detected in the videos such as preparing for action or coming to the queue. This detection may occur when player have a bigger activity measure than the threshold set, Fig. 12.

Fig. 12. Examples of actions outside of the scope

The errors may occur when the players are not detected or tracked correctly across the video. Figure 13 shows an incorrect example of shooting action that was caused by poor player detection.

Fig. 13. Examples of incorrect shooting action collages

Figure 14 shows a few examples in which the bounding box does not cover the full body of the player, and therefore some of the thumbnails on the collage lack the hands, head or legs of the player. This kind of problem should be reduced by additional training of the object detector network on examples from the handball domain. Due to incorrect detection, some of the thumbnails do not include players at all. Fortunately, although the scenario is complex, with artificial light and shadows that occurs very rarely.

Fig. 14. Examples of imprecise player detection

Another kind of error occurs in clotted scenes due to errors in tracking, e.g. when two players cross each other in the field and alternately cover each other and the trajectory tracking switches to the wrong player mid-sequence. In this case, the leading player may be correctly identified in only a part of the sequence, Fig. 15.

Fig. 15. Examples of problems in player tracking

Further training of the Mask R-CNN network on additional examples from the sports domain should prove useful because players performing sports actions exhibit greater variability of possible appearances than is expected for the usual "person" class. This is reflected in the results, as the player detection usually fails at exactly those frames of video where the player is at most extreme body position, e.g. the top of a jump-shot.

V. Conclusion

To build successful machine learning models, large amounts of labeled data of training examples is required. In this paper, we present an approach for building an action recognition learning dataset in the domain of sports videos that strives to minimize the tedious and time consuming manual labor required to label the individual players performing the chosen actions.

The method utilizes the existing deep learning object recognition methods for automatic player detection, and combines the obtained location information with a player activity measure based on spatio-temporal interest points to track active players, i.e. players that are performing the currently relevant action.

We have successfully applied the method to a challenging sports dataset of real-world handball practice videos where multiple players and bystanders appear simultaneously on the scene and can perform different actions. The videos were recorded indoors and outdoors in conditions that can be expected in amateur setting with consumer camera equipment and no special lighting. The proposed method has significantly simplified and sped-up the preparation of the action dataset, as only the action beginning and end times had to be manually defined. The players were detected and tracked automatically, and the player performing the desired action was correctly tracked and selected in 84% of all cases.

The used Mask R-CNN detector proved well suited for detection of players in the handball footage with few false positives, and with mostly good detections even when players were inside a group or away from camera, except in some cases of occlusion.

Still, we plan to train the Mask R-CNN network on additional examples from the sports domain to improve the performance with extreme positions.

Also, we plan to extend the method to handle changing player activity throughout longer video sequences, to reduce or remove the requirement for manual temporal segmentation of videos.

Likewise, the method will be extended to more complex actions that involve two or more players at the same time, like crossing or interaction with opponent defense players, to enable further expansion of the action recognition dataset with more data from realistic competition settings and matches.

References

[1] Lin, T, et al. (2014, September). Microsoft coco: Common objects in context. In European conference on computer vision (pp. 740-755). Springer, Cham.

[2] Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: a local SVM approach. In Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on (Vol. 3, pp. 32-36). IEEE.

[3] Blank, M., et al. (2005, October). Actions as space-time shapes. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on (Vol. 2, pp. 1395-1402). IEEE.

[4] Yuen, J., Russell, B., Liu, C., & Torralba, A. (2009, September). Labelme video: Building a video database with human annotations. In Computer Vision, 2009 IEEE 12th International Conference on (pp. 1451-1458). IEEE.

[5] Vondrick, C., Patterson, D., & Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation. International Journal of Computer Vision, 101(1), 184-204.

[6] All, K., Hasler, D., & Fleuret, F. (2011, June). FlowBoost—Appearance learning from sparsely annotated video. In Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on (pp. 1433-1440). IEEE.

[7] de Souza, C. R., Gaidon, A., Cabon, Y., & Pena, A. L. (2017, July). Procedural generation of videos to train deep action recognition networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Vol. 2).

[8] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 2980-2988.

[9] Ivan Laptev, On Space-time Interest Points, International Journal of Computer Vision 64(2/3), 107–123, 2005

[10] Chakraborty, B., Holte, M. B., Moeslund, T. B., & Gonzàlez, J. (2012). Selective spatio-temporal interest points. Computer Vision and Image Understanding, 116(3), 396-410

[11] S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, June 1, 2017.

[12] Munkres, J. (1957). Algorithms for the assignment and transportation problems. Journal of the society for industrial and applied mathematics, 5(1), 32-38.

[13] S.M. Smith and J.M. Brady. ASSET-2: Real-time motion segmentation and shape tracking. IEEE-PAMI, 17(8):814–820, 1995.

[14] D.G. Lowe. Object recognition from local scale-invariant features. In Proc. ICCV pages 1150–1157, Corfu, Greece, 1999.

[15] M. Burić, M. Pobar, M. Ivašić-Kos, "Object Detection in Sports Videos," *2018 MIPRO*, Opatija, 2018.

[16] M. Ivašić-Kos, M. Pobar, S. Slobodan. Two-tier image annotation model based on a multi-label classifier and fuzzy-knowledge representation scheme. Pattern recognition. 52 (2016); 287-305

View publication stats