Detection of the leading player in handball scenes using Mask R-CNN and STIPS

Pobar, Miran; Ivašić-Kos, Marina

Source / Izvornik: **Proc. SPIE 11041, Eleventh International Conference on Machine Vision (ICMV 2018), 2019, 11041, 501 - 508**

Conference paper / Rad u zborniku

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

https://doi.org/10.1117/12.2522668

Permanent link / Trajna poveznica: https://urn.nsk.hr/urn:nbn:hr:195:714143

Detection of the leading player in handball scenes using Mask R-CNN and STIPS

M. Pobar, M. Ivašić-Kos

University of Rijeka, Department of Informatics, Rijeka, Croatia
marinai@inf.uniri.hr, mpobar@inf.uniri.hr

ABSTRACT

In team sports scenes, recorded during training and lessons, it is common to have many players on the court, each with his own ball performing different actions. Our goal is to detect all players in the handball court and determine the leading player who performs the given handball technique such as a shooting at the goal, catching a ball or dribbling. This is a very challenging task for which, apart from an accurate object detector that is able to deal with cluttered scenes with many objects, partially occluded and with bad illumination, additional information is needed to determine the leading player. Therefore, we propose a leading player detector method combining the Mask R-CNN object detector and spatio-temporal interest points, referred to as MR-CNN+STIPs. The performance of the proposed leading player detector is evaluated on a custom sports video dataset acquired during handball training lessons. The performance of the detector in different conditions will be discussed.

Keywords: object detectors; handball sports scenes; Mask R-CNN, spatio-temporal interest point – STIP, tracking, attention

1. INTRODUCTION

Object detection is a prerequisite for many applications of computer vision such as action recognition, security, surveillance, image retrieval, autonomous driving cars, etc. The task of object detection is to find instances of real-world objects in images or videos such as people, cars, faces etc. To detect an object, it is necessary to determine the location of that object and to predict the class to which it belongs. The challenge is to solve both object location and object classification problems, so the choice of the right object detection method depends on the problem that needs to be solved.

In our case, object detection is a precondition for action recognition on handball scenes. Handball is a very fast and dynamic team sport and the shape and appearance of a player can change fast and may vary greatly in time. Also, the motion blur and shadows that players cast under artificial illumination are often present in the videos, making the problem even harder.

The handball game has its own rules but in order to keep students motivated during handball lessons, teachers modify the rules of the game to maintain a high level of activity for students while developing and practicing handball techniques. The aim is to organize a lesson such that the students have the longest possible active time with variable and tactical exercises that mimic real game situations. For this reason, most activities are performed in parallel to keep the waiting time between the activities as short as possible and to make more repetitions.

Each student usually has his own ball, so that several activities take place at the same time. For example, when practicing the throwing techniques, there is one student who shoots the ball at the goal, the goalkeeper that moves to save the goal, while others who have performed this activity collect their balls around the court and run to their position in the queue.

A player can shoot the ball, dribble it, or pass it to a teammate. The students should practice various throwing techniques whether with jump or standing, different fakes and moves to surprise the opponent and win the game. Also, there are special techniques that depend on the court position of the player such as goalkeeper, center, wing, pivot or back-players as well as techniques that are related to offense or defense playing ability and handball skills.

The object detector to be successful should thus be able to deal with challenging conditions like a variable number of players, different player positions, a wide range of possible player sizes ranging from those that can cover most of the image to the one that is far away from the observer or is occluded. Also, some very small objects of just a few pixels such as a ball can have significant information for the interpretation of a scenes or for action recognition. [1] The background is often cluttered with challenging indoor illumination.

Our goal is to detect players on the court and determine the player performing the given handball technique. Here we call it the leading player.

To this end, we propose a leading player detection method that combines an object detection and tracking step with a spatio-temporal interest detection step. Recent state-of-the-art object detection methods rely on deep convolutional neural networks (CNNs), which have achieved a remarkable increase in the accuracy of object detection in images. For this reason, for the object detection step, we will use the Mask R-CNN network (Mask R-CNN) [2] that has proven to be successful in classification and detection tasks in real-world images.

To discern the activity level and thus the interest of the detected players, we use the space-time interest points (STIP) that jointly embed space and time information detected within the player bounding boxes. In this paper, we referred to the proposed method as MR-CNN+STIPs.

The rest of the paper is organized as follows: in Section II. we will briefly review the object and space-time interest point detection methods. In Section III., the proposed method that combines the Mask R-CNN object detector and spatio-temporal interest points to determine the leading player in sports scenes, is described. We have applied the proposed

method on a custom dataset consisting of indoor and outdoor handball scenes recorded during handball school. The comparison of the detector performance in different conditions is given in Section IV. The paper ends with a conclusion and the proposal for future research.

2. OBJECT AND SPACE-TIME INTEREST POINT DETECTORS

2.1 Object detectors

The goal of object detection and recognition is to locate and classify objects on the scene. The detected objects are typically marked with a bounding box and labeled with their corresponding class labels.

In object recognition, the current focus is on convolutional neural networks (CNNs), influenced by the results of AlexNet [3]. Subsequent CNN-based systems like VGGNet [4], Inception [5], etc. were at first only used for classification, i.e. to determine the class of object of interest in the image, but later have been extended to detect and localize objects as well. This is commonly achieved by independently processing parts of images that are isolated from the whole image by positioning rectangular windows over overlapping parts of the image. To account for different possible object sizes, multiple window scales are used. If a classifier recognizes the contents of a window as an object, it is labeled with the appropriate class label and the corresponding window is used as the object bounding box. The result of processing the whole image is a set of bounding boxes and corresponding class labels. Since the window is sliding over overlapping image areas, a large number of duplicated predictions can be generated, which can then be discarded using a confidence threshold. Since the sliding window approach in essence performs the image classification for each window position, the naïve implementation can be much slower and computationally expensive in comparison with the simple classification.

An optimized version of the sliding window approach was used in the Region with CNN features (R-CNN) network (Fig. 1) [7]. Here, instead of considering all positions of the sliding window, only a subset, here called region proposals generated by selective search process [8] is further processed by the CNN to extract features. The classification is then performed using support vector machines, and finally the bounding box coordinates are tightened using linear regression considering the determined object class.

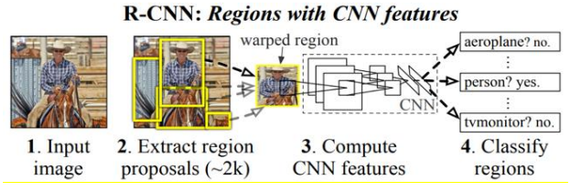

Figure 1. The principle of selective search inside R-CNN according to [7]

A series of methods [2, 9, 10] further improved and refined this approach. In Fast R-CNN [9], the feature extraction is sped up by performing the forward pass through the network only once on the whole image, and by selecting the features for each region from the resulting feature map of the whole image. Also, classification was performed within the same CNN framework, instead of using SVMs. In Faster R-CNN [10] the region proposal has been integrated within the same CNN architecture as the first stage of the network, called region proposal network (RPN), replacing the external selective search process. The RPN is a deep fully convolutional network that takes an image and outputs a feature map, upon which a smaller network is applied in a sliding window fashion. The smaller network takes a spatial window of the feature map, reduces the feature dimension, and then feeds them to the two fully-connected layers that output in parallel the bounding box coordinates of proposed regions and an „objectness“ score for each box.

The Mask R-CNN [2] adds a third parallel output to the Faster R-CNN, using a small fully connected network to predict the segmentation masks on each region of interest. Masks and bounding boxes are generated in Mask R-CNN independently from classification, and the relevant masks and bounding boxes are finally selected using the result of the classification branch.

The Mask R-CNN proved to be appropriate for player detection in the footages of team sports because it can successfully detect individual players even when they are inside a group, further away from the camera [11].

2.2 STIPS detectors

Many interesting events and actions in sports videos are characterized by strong variations in velocity and appearance over time. More generally, points in videos or local image structure with non-constant motion might correspond to the moving objects in the real world and therefore can contain important information in both the spatial and the temporal dimensions about the action and changes in the environment. Points in the spatial domain with a significant local variation of image intensities, denoted as “interest points” have been extensively investigated in the past and successfully applied in a number of applications such as optic flow estimation and tracking [12], image indexing and recognition [13].

For jointly embedding space and time information, the idea of space 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”, (STIP). STIPs are extracted encoding the temporal information directly into the local salient feature. To differentiate events from each other and from noise, one approach is to compare their local neighborhoods and assign points with similar neighborhoods to the same class of events [12].

Similar popular solutions of STIP detectors are based on the detection of spatio-temporal corners derived from Harris corner operator (Harris3D) proposed by [14], Dollar’s detector [14] that uses a Gaussian filter in space and a Gabor band-pass filter in time and obtains a denser sampling by avoiding scale selection and Hessian3D derived from SURF [16]. Widely used descriptors that represent the dynamic content of the cuboid volume are histogram of 3D gradient orientations

(HOG) based on space-time pixel values derivatives that represents the appearance and histogram of Optical Flow magnitude and orientation (HOF) that models the motion such as HOG/HOF [14], Dollar [15], HoG3D [16], Extended SURF [17].

3. LEADING PLAYER DETECTION METHOD MR-CNN+STIP

In handball, multiple players are present in the scene at the same time, and although they all might move and interact, not all players contribute to the currently most relevant action. The goal of this experiment was to automatically determine the player or players in the scene that are responsible for the current action.

An overview of the proposed method is shown in Fig. 2. For each input video frame, STIPs and player bounding boxes are detected and the bounding box with the highest density of contained STIPs is marked. The detected bounding boxes are tracked across the whole sequence to form player trajectories. The trajectory that has the most marked bounding boxes in the whole video is considered to belong to the active, leading player.

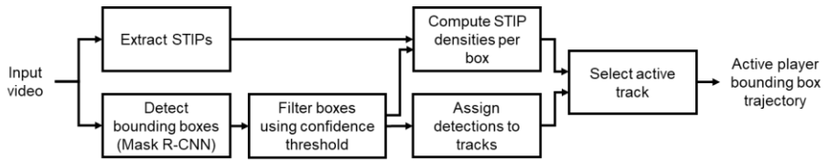

Figure 2. An overview of the active player track detection.

3.1 Detection of players

To detect the players in the scene, the Mask R-CNN detector from the Detectron suite [18] was used. The detector was in the standard Resnet-101-FPN network configuration and used with the parameters pre-trained on the COCO dataset, with no additional training on our dataset. The Mask R-CNN network outputs object masks for the detected objects, their bounding boxes, and the corresponding class labels and confidence values. In this experiment, only the bounding boxes and confidence values for the objects of the class "person" were used.

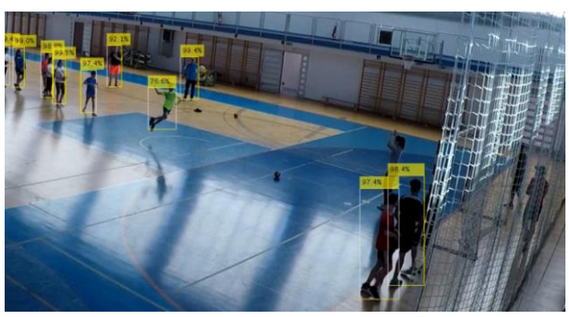

Figure 3. Bounding boxes with confidence values as results of Mask R-CNN person detection

3.2 Player tracking

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

To eliminate false detections, only the bounding boxes with confidence scores higher than a threshold value were considered in this step. The threshold value was experimentally set to 0.6, which provided a good balance of high detection rate with few false positives.

Initially, in the first video frame, a track id is assigned to each detected bounding box.

Figure 4. In each video frame, a bounding box has its track ID

Then, for each next frame, the assignment of the individual detected bounding boxes to tracks is done using the Munkres' version of the Hungarian algorithm which minimizes the total cost of assigning the detections to tracks. The cost of assigning every detection to each track considers both 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 precisely, a new bounding box is determined to correspond to the minimal cost computed as a weighted sum of the Euclidean distance between the detected bounding box centroids (C_b) and the predicted track centroids ($C_{b+1}{}'$) and of the absolute area difference of the detected boxes area (P_b) and the last box area assigned to tracks (P_{b-1}) (1):

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

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, 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, it has shown uneven performance and was not used further in the experiment.

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. A cost threshold of T is used to control the maximum allowed distance between a detection 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 consecutive frames, no new detections are added to the track. The values of M and T are experimentally set to values of 20 and 100.

3.3 Determining the leading player

The bounding boxes capture the location of the players but don't carry information about the movements of the players. On the other hand, the STIPs are local features that capture the spatio-temporal „interestingness" at different points in image that can correspond to any object or to the background. Therefore, the idea is to merge two different kinds of information in order to detect an active, leading player that is performing action related to handball techniques among others that have already performed desired action and are preparing for next repetition of that action or for 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 capture the two pieces of information, for each detected bounding box (B) in each frame, an activity measure (A_b) is computed in parallel, using the density of spatio-temporal interest points (STIPs) in the box (2):

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

where P_b is the area of the box B.

Figure 5. Fusion of bounding box and spatio-temporal interest points

The STIPs are extracted using the selective STIPs method [19] with default parameters from the whole video. The bounding box containing the highest density of STIPs is considered to correspond to the player with the highest activity and is marked. The detected bounding boxes are tracked across the whole sequence to form player trajectories.

Figure 6. Tracking the leading player (marked with thick white bounding box)

The tracking id of the bounding box with the highest activity in each frame is recorded in a vector of active tracks. Finally, the track whose id appears most often in the vector of active tracks is selected as the active player's track (Fig. 8). The trajectory that has the most marked bounding boxes in the whole video is considered to belong to the active player.

Figure 7. Detected leading player (white box) and her trajectory through the whole sequence (yellow line)

4. EVALUATION OF LEADING PLAYER DETECTION PERFORMANCE AND DISCUSSION

We have tested the proposed method on a custom dataset consisting of indoor and outdoor sports footage during practice and competition, recorded during the handball school. A handball game for students was organized 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.

The dataset consists of 751 videos of 7 different action classes. In each video multiple players appear (10 in average) and each of them can perform an action. However, each file is labeled only with a single action of interest, performed by the leading player. The total duration of the recorded material is 1990s. The scenes were captured using stationary GoPro cameras from different angles and in different lighting conditions (indoor and outdoor). The cameras in indoor scenes were mounted at a height of around 3.5 m to the left or right side of the playground. The outdoor scenes have the camera at a height of 1.5 m. The videos are recorded in full HD resolution (1920x1080) and with 30 frames per second. Both object detection with Mask R-CNN and STIPs extraction was performed on full resolution videos and no frame skip.

First, just the used Mask R-CNN object detector was tested on the dataset in isolation. The detectors performance was evaluated in terms of recall, precision, and F1 score [20], 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. Fig, 9. shows the results of the evaluation in the case of a simple and complex scenario. A simple scenario includes fewer objects, up to 8, close to the camera. A complex scenario is when the number of objects on the scene is equal and greater than 9, away from the camera and with the occlusions.

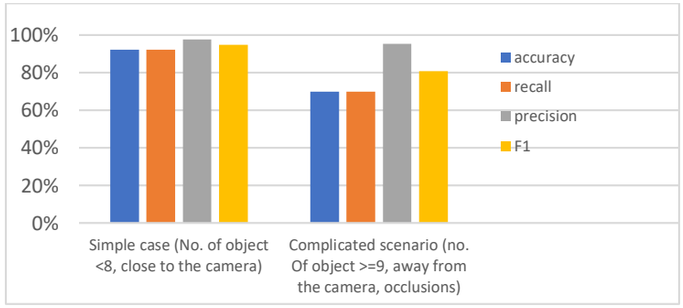

Figure 8. Evaluation results of player detection with Mask R-CNN in simple and complex scenarios.

The average number of detected objects is 26.67±3.20, and the average number of those whose confidence is greater than 85% is 10.21±0.95.

Next, the whole leading player detection was tested. Since the proposed method always selects one player as the leading one, we used the true positive rate to quantify its performance. We count as a true positive only those cases when the leading player is correctly detected throughout the whole sequence (fig. 10, left). The true positive rate is the number of true positives divided by the number of tested videos/sequences.

The errors may occur when the players are tracked correctly across the video, but the wrong player is selected as the leading one. This may occur either because the leading player was not detected at all, or because a different player than the leading one had a bigger activity measure, fig. 10, right. Another kind of error occurs due to errors in tracking, e.g. when two players cross each other in the field 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.

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

The results of the evaluation are presented in Table 1.

Table 1. True positive rates for leading player detection.

Action	Passing	Shooting	Jump shot	Dribbling	Average
TPR %	82.35	89.65	81.11	83.33	84.11

The best performance was achieved with the shooting action, where players shoot the ball from the ground, while the somewhat similar action, shooting from the air mid-jump proved more difficult. This can be in part attributed to the different performance of the Mask R-CNN object detector for these two cases. In the case of the more dynamic jump shot, the player appearance is often very different than the typical relaxed stance of persons that the network was probably trained on, so the lead players were more often not detected at all in the case of the jump shot. In the case of ground shot, players silhouette is much closer to an appearance to a pedestrian. In addition, in the jump shot case, even when detected, the player's bounding box more often covered only a part of the body (e.g. fig. 6 right). This made the activity assessment less precise, since a greater number of STIPs related to most active body parts (arms, legs) fell out of the bounding box, reducing the activity score for that player. In that case, some other player was more often wrongly selected as active. This problem should be reduced by additional training of the object detector network on examples from the sports domain.

The passing action is also more complex since it involves two players, one who passes the ball, and the one who catches the ball. Since both players may show similar measures of activity, the number of frames in which the one or the other is selected as leading may be similar, leading to less reliable detection.

5. CONCLUSION

In the majority of acquired video footages, Mask R-CNN has proven to be successful, with a few false detections except in case of occlusion. However, due to a very large range of possible appearances of players performing sports actions, to improve the detection of players and bounding boxes, the Mask R-CNN network should be trained on additional examples from the sports domain.

The proposed method of detecting the leading player in team sports proved successful in typical handball training scenarios. The results are good enough to be used as a means to automatically generate ground truth labels for an action recognition database. In that case, instead of manual annotation of players, only a manual verification is required, greatly reducing necessary time and effort.

In the future, the method should be extended to better handle the case of actions involving multiple players, e.g. crossing, where a single leading player does not completely represent the whole action.

Also, the detector should be trained to detect small objects such as a ball that carries a lot of information useful for prediction of actions and is the center of attention in handball game.

ACKNOWLEDGMENT

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).

REFERENCES

[1] M. Burić, M. Pobar, M. Ivašić-Kos, "An overview of action recognition in videos," *2017 MIPRO*, Opatija, 2017, pp. 1098-1103.

[2] 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.

[3] A. Krizhevsky, I. Sutskever and G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks, Advances in Neural Information Processing Systems, 2012., pp. 1097-1105.

[4] K. Simonyan, A. Zisserman. "Very Deep Convolutional Networks For Large-Scale Image Recognition," arXiv:1409.1556., 2014.

[5] C. Szegedy et al., "Going deeper with convolutions," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, 2015, pp. 1-9.

[6] M. Pobar, M. Ivašić-Kos, Multi-label Poster Classification into Genres Using Different Problem Transformation Methods; Computer Analysis of Images and Patterns, CAIP 2017, Lecture Notes in Computer Science, vol. 1042

[7] Girshick, J. Donahue, T. Darrell, J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, 2014, pp. 580-587.

[8] J.R.R. Uijlings, K.E.A. van de Sande, T. Gevers, and A.W.M. Smeulders, „Selective Search for Object Recognition", International Journal of computer vision, 104(2), 154-171.

[9] R. Girshick, "Fast R-CNN," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, 2015, pp. 1440-1448.

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

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

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

[13] D.G. Lowe. Object recognition from local scale-invariant features. ICCV p. 1150–1157, Corfu, Greece, 1999.

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

[15] P. Dollár, V. Rabaud, G. Cottrell, S. Belongie, Behavior Recognition via Sparse Spatio-Temporal Features ICCV VS-PETS 2005, Beijing, China.

[16] A. Klaser, M. Marszalek, C. Schmid. A Spatio-Temporal Descriptor Based on 3DGradients. BMVC 2008 - 19th British Machine Vision Conference, Sep 2008, Leeds, United Kingdom. British Machine Vision Association, pp.275:1-10, 2008

[17] Willems, G., Tuytelaars, T., Gool, L.: An e_cient dense and scale-invariant spatiotemporal interest point detector. In: ECCV. (2008) 650-663

[18] Detectron suite, https://github.com/facebookresearch/Detectron/blob/master/CONTRIBUTING.md

[19] 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

[20] M. Ivašić-Kos, M. Pobar, Multi-label Classification of Movie Posters into Genres with Rakel Ensemble Method; Artificial Intelligence XXXIV. SGAI 2017. Lecture Notes in Computer Science, vol. 10630; Cambridge: Springer, 2017. 370-383

AUTHORS' BACKGROUND

Your Name	Title*	Research Field	Personal website
Miran Pobar	postdoctoral fellow	Pattern recognition, Machine learning, Computer vision	https://portal.uniri.hr/portfelj/755
Marina Ivašić-Kos	assistant professor; senior research associate;	Pattern recognition, Machine learning, Computer vision	https://www.linkedin.com/in/marinaivasickos/ https://portal.uniri.hr/portfelj/960