Multiple object tracking for video-based sports analysis

Julius Gudauskas and Žygimantas Matusevičius

Kaunas University of Technology, Studentų g. 50, Kaunas, Lithuania

Abstract

Multiple object tracking (MOT) is a challenging task in computer vision. Many algorithms have been proposed to track multiple targets for video surveillance, team-sport analysis, or human–computer interaction. Recent studies have already indicated that multiple object tracking could provide valuable information in team sports analysis. Therefore, in this paper, we investigate object tracking techniques for paralympic team sport – goalball. Different tracking methods have been implemented and compared, evaluating prediction accuracy and performance speed in players and the ball tracking.

Keywords

Multiple object tracking, MOT, SOT, CNN, ONNX, Goalball, Boosting, CSR-DCF, KCF, MOSSE, TLD.

1 Introduction

As computer computing capabilities increase, image processing technologies are becoming increasingly important. Image processing involves many processes, but the main goal is to detect objects and identify their movement's nature. Object recognition and tracking of its dynamics serve in many areas of life, such as robotics [1], smart cities [2], [3], medicine [4], [5], or human-computer interface [6]. Besides those already mentioned herein, we would like to address further how better to integrate computer vision technologies into paralympic sports applications. Recently, there has been a growing interest in developing intelligent systems for detecting and tracking player's movements to analyse sports players' performance during games or training sessions and improve their performance. The information derived from such analysis is valuable to sports experts and coaches since it allows them to understand major mistakes better, identify trainees' weaknesses, and modify training and strategic plans accordingly. Although the number of sports innovations has been increasing and technological solutions are being implemented even by archaic sports, for example, video assistant referee (VAR) system in football, sports for the disabled remain a side-line - for instance, goalball one of the most popular sports among the blind. In order to raise or at least maintain the competition in this sport, it is necessary to implement novel technological solutions in the process of training and tournaments. Artificial intelligent based technologies should be integrated in order to acquire multidimensional information. According to the experts, the essential task is to monitor the ball's and players' movements. However, tracking ball and players over a large playing area is a challenging problem for several reasons. First, the players move quickly and have large variations in their silhouettes. Goalball is a team sport; therefore, multiple player tracking must be performed, but it may be complicated when players are spatially close together. Second, the dynamic nature of ball appearance, movement, and continuously changing background make the detection and tracking processes even more challenging [7]. Besides, the ball size is relatively small compared to other objects in a frame, and it can be overlapped or covered by other objects. This work aims to integrate image processing techniques into goalball game video analysis for real-time detection and tracking of multi-players and the ball. Different tracking methods have been implemented and compared in terms of detection precision and speed.

2 Overview

Some object tracking strategies have been implemented, but the best solution has not yet been found. Next, we examine the existing object tracking techniques and object tracking solutions in sports.

2.1 Related works

SAP develops solutions for video-based sports analytics. For the football world championship 2014 in Brazil, SAP with German Football Association successfully developed Match Insights analytical solution. It was decided to integrate it with Panasonic video and tracking software [8] to improve the solution.

VisualiZation in real-time (Vizrt) provides content creation, control, and delivery tools for the digital media business. The company's products include software for designing real-time 3D graphics and maps, envisioning sports analyses, controlling media assets, and obtaining single workflow solutions for the digital broadcast trade [9], [10].

PITCHf/x data set is a free source granted by Major League Baseball Advanced Media (MLBAM) and Sportvision. Brooks Baseball [11] performs methodical innovations to this data to increase its worth and usability. They manually analyze the Pitch Info by using many parameters of each pitch's trajectory and approve the parameters against some other sources such as video proof and direct interaction with on-field personnel (e.g., pitching coaches, catchers, and the pitchers themselves). The trajectory data's default values are somewhat altered to align them more nearly with the actual values.

Sportradar [11], a Swiss corporation, concentrates on accumulating and examining data related to sports results by cooperating with bookmakers, widespread football associations, and global football associations. Their performing projects include collecting, processing, monitoring, and selling sports data, appearing in a different collection of sports-related live data and digital content.

2.2 Multiple object tracking based on single object tracking

Multiple object tracking (MOT) is one of the most challenging tasks in computer vision. A reliable and universal solution to this problem is not yet known - often, several objects are tracked using a single object tracking (SOT) method. With this tracking method, each object is tracked separately and independently of the other objects. The article [12] proposed a powerful real-time tracking method Boosting, that considers the tracking problem as a binary classification problem between object and background. Most existing approaches build a representation of the targeted object before the tracking function begins and therefore utilize an established representation to handle appearance adjustments during tracking. However, this method does both - adjusting to the variations in appearance during tracking and selecting suitable features that can learn any object and discriminate it from the surrounding background. In Discriminative Correlation Filter with Channel and Spatial Reliability (CSR-DCF), the reliability map adapts the filter support to the object suitable for tracking, overcoming both the circular shift problems and enabling an arbitrary search range and the rectangular shape assumption's limitations [13]. The CSR-DCF has the highest performance on standard benchmarks – OTB100, VOT2015, and VOT2016 while running in real-time on a single CPU. Despite using basic features like histogram of oriented gradient (HOG) and Colornames, the CSR-DCF performs parallel with trackers that apply computationally complex deep Convolutional Networks but is noticeably faster. In [14], originators demonstrated that it is possible to analytically model natural image translations, showing that the resulting info and kernel matrices become circulant under some conditions. The DFT's diagonalization presents a general blueprint called Kernelized Correlation Filter (KCF) for creating fast algorithms that deal with translations. This blueprint has been applied to linear and kernel ridge regression, obtaining the highest development trackers that work at hundreds of FPS and can be implemented with a few code lines. The visual tracking problem, which is traditionally solved using heavyweight classifiers, complex appearance models, and stochastic search methods, can be replaced by effective and more straightforward Minimum Output Sum of Squared Error (MOSSE) correlation filters [15]. However, there are several ways how this tracker can be improved. For example, if the

target's appearance is relatively steady, drifting could be eased by occasionally recentring the filter based on the initial frame. Also, the tracker can be extended to estimate scale and rotation changes by filtering the tracking window's log-polar transform after an update. In paper [7], authors studied the problem of tracking an object in a video stream, where the object changes appearance frequently moving in and out of the camera view. They designed a new Tracking, Learning, and Detection (TDL) framework. Many challenges have to be addressed to get a more trustworthy and general system based on TLD. For example, TLD does not perform well in the case of full out-of-plane rotation. In that case, the Median-Flow tracker drifts away from the target and can be re-initialized if the object comes back with an appearance seen/learned before. The current implementation of TLD trains only the detector, and the tracker stays fixed. As a result, the tracker always makes identical errors, and currently, it tracks a single object. Multi-target tracking opens engrossing questions about how to train the models and share features to scale jointly.

2.3 Multiple object tracking based on object detection and position forecasting

We can also rely on recognition-based solutions to solve the problem of tracking multiple moving objects. These algorithms' idea is to detect the tracked objects in each analyzed frame and classify them into sets of moving objects. This problem is usually framed as a data linking task, but several obstacles can lead to poor tracking accuracy. To identify tracked objects in a frame can be applied various neural network-based or non-neural network-based algorithms. Such classical methods as the Viola-Jones algorithm work in real-time by analyzing the image's pixels [16]. Although the algorithm is quite primitive, it has pretty high accuracy and real-time speed. This algorithm can be taught to detect classes of different objects (applied to different subtasks such as pedestrian and car), but due to the algorithm's favourable properties, this algorithm is usually applied in face recognition. [18] The object detection process can be established using HOG [14], scale-invariant feature transformation [16], Haar cascade classifiers [16], etc. These algorithms are used to determine low-level feature information. More complex tasks usually require obtaining higher-level information, and that is possible using deep learning techniques. A convolutional neural network (CNN) a class of deep neural networks, most commonly applied for image recognition tasks [16]. You Only Look Once (YOLO) is a deep learning algorithm for object detection, which is most fast and accurate than most other algorithms [16]. By dividing the input image into areas and predicting the boundary box's coordinates and the class's probability for each region, it converts object detection problems into regression issues to achieve end-to-end detection. YOLO can work well for multiple objects where each object is associated with one grid cell. However, in the case of overlap, in which one grid cell contains two different objects' centre points, we can use anchor boxes to allow one grid cell to detect multiple objects. The common challenge complicates the multiple objects tracking and detrimental to the result – frequent occlusions, similar appearance, interactions between multiple tracked objects, the unstable appearance of the object in the video, etc.

3 Proposal

In the following part of the article, we will provide proposals for multiple object tracking. The presented algorithms are designed to solve the players tracking problem in targeted video.

3.1 MOT using SOT

Firstly, multiple object tracking (MOT) is developed by employing independent single object trackers (SOT). MOT model has a list of tracked objects, and each of objects has its *tracker*, *id*, and *rectangle object*, which stores the metrics of the tracked object: *x* and *y* coordinates, height and width of the region of interest (ROI) (see **Figure 1**).

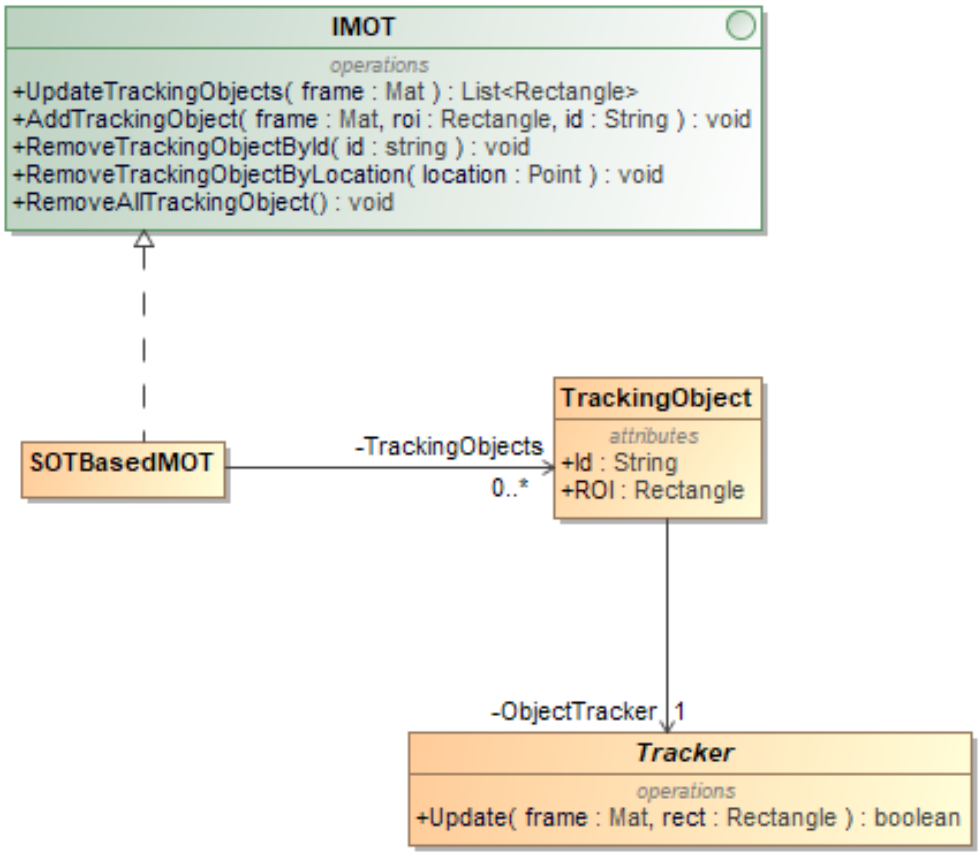

Figure 1: SOT based MOT model

The process of object tracking using Unified Modeling Language (UML) notation is provided in **Figure 2**. First, a video file is selected, and MOT initialization is performed. After the initialization of the model, the objects to be tracked are marked. This process is performed manually. Finally, it is possible to start processing video frames, where each frame is used to update the MOT model.

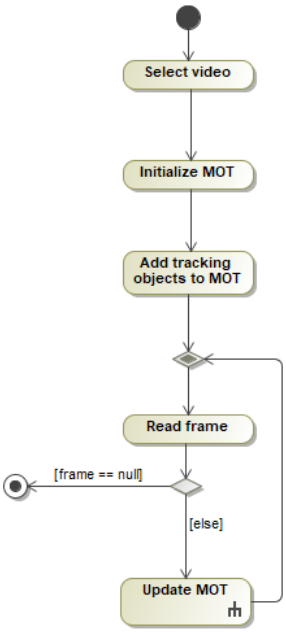

Figure 2: SOT based MOT process

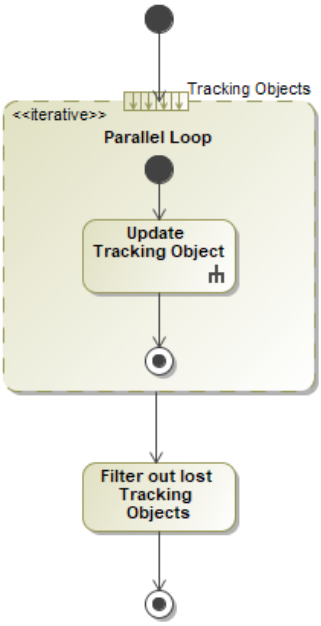

Figure 3: Update MOT process

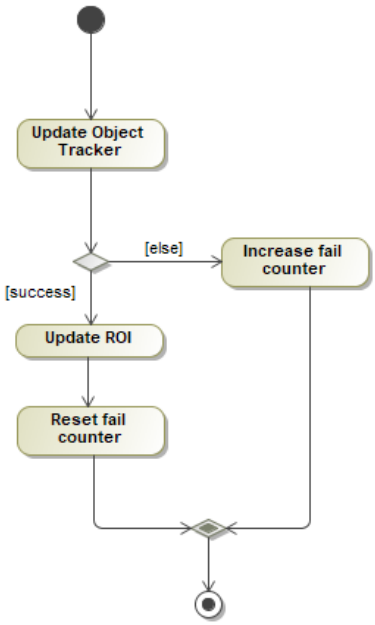

Figure 4: Update tracking object process

During the MOT model update (see **Figure 3**), the frame is used to update each tracked object. Because objects are tracked entirely independently, this process can be parallelized. After updating tracked objects, MOT removes from the list those objects that have not been successfully updated for a certain period of time - it is assumed that the object has been lost. During the tracked object update process (see **Figure 4**), the object tracker is updated. If this operation is performed successfully, then the rectangle object is updated, and the failure counter is restored; otherwise, the fail counter increased.

3.2 MOT using CNN object detection

An alternative way to track multiple moving objects is to use constant object detection and detected object classification. To solve this problem, a method of object recognition that provides the highest possible accuracy, as well as a method of classifying objects according to the previous coordinates of the presence of each moving object, is required. Convolutional Neural Network (CNN) allows forming a multi-layered model, which can provide an advantage in analyzing more than one feature without compromising speed. For CNN model training, 117 different shots of a goalball match have been used. All the training data has been marked with the bounding box required for prediction (see **Figure 5**).

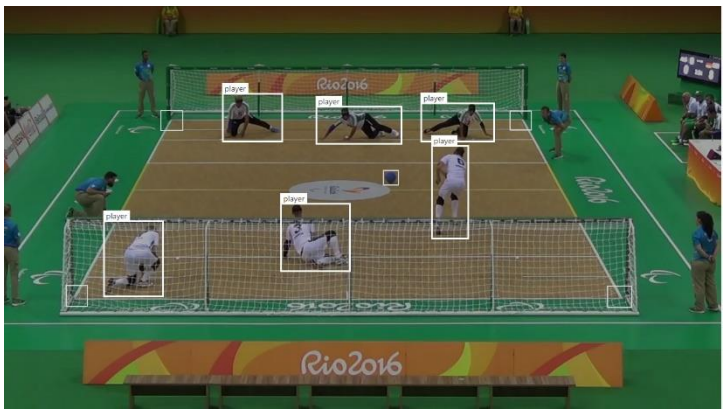

Figure 5: Preparing a frame for training

To evaluate the performance of the CNN model, three accuracy characteristics have been used: precision, recall, and the mean average precision (mAP):

The trained model can be exported and applied locally. Depending on the technology used, the model format can also be selected in different ways. The Open Neural Network Exchange (ONNX) format model is used for this study. ONNX provides definitions of an extensible computation graph model, built-in operators, and standard data types focused on inferencing (evaluation). The model was constructed using eight layers with the input image in BGR format. Trained CNN model provides composed of bounding boxes, class labels, and confidence levels (see **Figure 6**). Each player is detected multiple times with a different probability. To remove unwanted redundancies, a filter is used that leaves only the bounding box satisfying the marginal probability. Recognition of players is not enough to track them. There is a classification problem in how to assign a bounding box to a particular player.

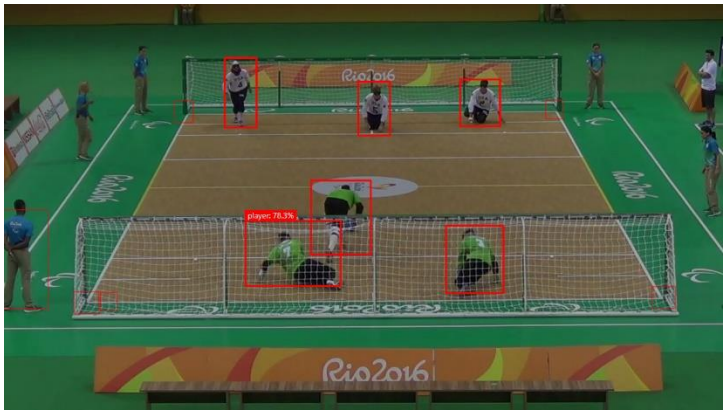

Figure 6: CNN predictions results

At the beginning of the analysis, the players being followed are marked. Having the start coordinate of each player, we can go through all the CNN results and assign each player the best bounding box. This step is repeated for each iteration of the refinement. Once the player was not detected using CNN (it usually appears when the player intersects), we keep the old coordinate and move to the next frame.

4 Experiment

The experiment was done using five different single object tracking methods: Boosting, CSR-DCF, KCF, MOSSE, TLD, and one multiple object tracking using CNN object detection. For testing, it was used three different goalball videos up to 1 minute long each. The experiment goal is to track six different players on the playfield marked from 1 to 6 (see **Figure 7**).

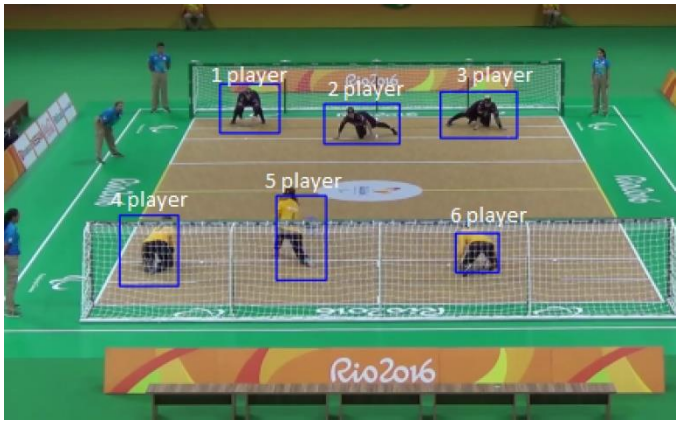

Figure 7: Tracked objects labels

The two most essential parameters in the evaluation of algorithms are:

- How accurately the player is tracked;
- How quickly the video is analyzed.

To evaluate player tracking accuracy, the number of frames in which a player was detected and tracked was processed. Because videos of specific durations and frames per second ratio were used for testing, we can evaluate each player's accuracy being tracked. However, not all algorithms can determine whether a player's position has been adjusted or not, so this metric is only valid for specific algorithms.

We have analyzed results of the KCF algorithm at the tracking in more details (see **Figure 8**). The algorithm provides the best tracking results for the third and the sixth players (in some cases, the first). The best results were obtained using the "Male BRA vs. SWE" video stream: the total accuracy of all tracked players is 89%. From the results, it cannot be said that the algorithm gives a stable and similar result under all conditions because it depends on many factors: the noise in the video material, bystanders, the exchange of players, the angle of the frames. It can be justified comparing the "Male BRA vs. SWE" and "Male LTU vs. USA" videos where a clear difference is visible. Using the same algorithm, total accuracy dropped from 89% to 81%. Supervising the algorithm revealed that players overlap more often in the less accurate video than in the greater accuracy provided one. Also, the tracked player is more often abandoned when making very sudden movements.

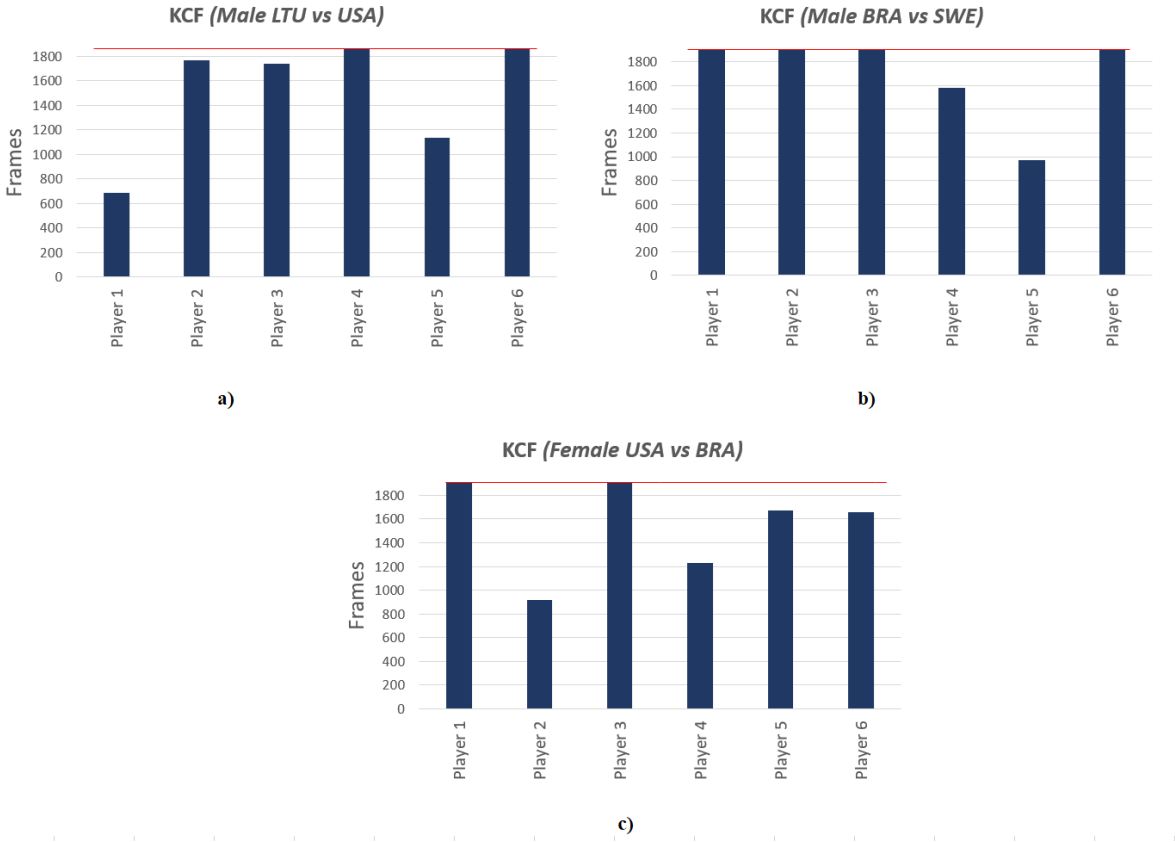

Figure 8: KFC algorithm results with different video streams

Another critical factor in the evaluation of algorithms is speed. Each algorithm is based on a different computational strategy, in which the speed may depend on different factors. After performing an experiment with each tracking algorithm and analyzing three videos for testing (see **Figure 9**), it was noticed that the MOSSE algorithm copes with the task even several times faster than the other algorithms. The slowest algorithm that the experiment was done is TLD.

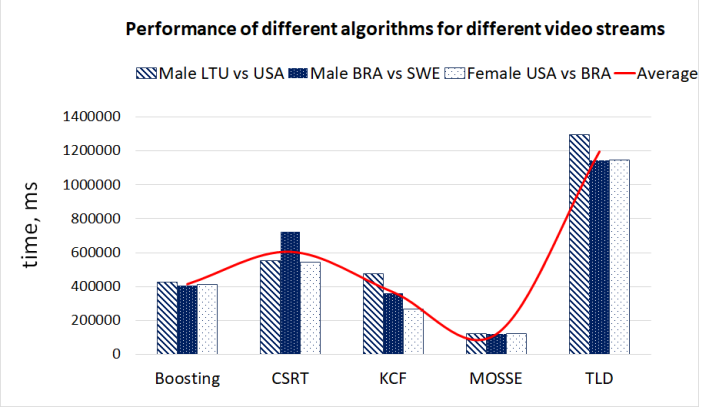

Figure 9: Algorithms performance

Another experiment - to use CNN for object detection and tracking. The model was trained using Microsoft Azure Custom Vision Service. The training process was performed using 117 different shots from the goalball game videos. In each frame, the players on the playing field and the ball were marked. The obtained results after training the convolutional neural network are provided in the **Table 1**.

Table 1
Model training results

Parameter	Value	Explanation
Over all precision of tags	96.9%	It measures how many of the predictions that the model made were actually correct
Over all recall of tags	82.4%	It measures how well the model can find all the positive predicted boxes
Over all Mean Average Precision (mAP)	66.8%	It is calculated by taking the mean average precision over all classes and overall IoU thresholds, depending on different detection challenges that exists

From the results, we can conclude that the model quite accurately predicted the players in the frames. Of these guesses, an average of 82% is accurate bounding boxes belonging to the hypothesized object. When it comes to recognizing the ball in the frame, the model performs worse. While the average is 89% of the shots, the ball is guessed; only an average of 39% is guessed in the proper right place. This may be because the ball is relatively small in the frame, and it is partially blocked by the players and sometimes merges with the background.

Additional experiments have been carried out to evaluate how the model performs in real test cases using video stream. The main task of the tracking algorithm is to solve the predicted bounding box classification. The prediction was made using CNN, and bounding box classification was performed using an algorithm that depends on previously detected player coordinates. The same data set, including three video streams, has been used in the experiment. Tracker provides quite a stable accuracy for each video stream (see **Table 2**).

Table 2
CNN based object tracking average results with different video streams

Video	Average accuracy
Paralympic Games 2016 Goalball Male LTU vs USA	88.12%
Paralympic Games 2016 Goalball Male BRA vs SWE	89.96%
Paralympic Games 2016 Goalball Female USA vs BRA	90.26%

Since the result determines in which part of the frame each player was detected and classified (see **Figure 10**), it is difficult to evaluate whether the classifier classifies the players with their belonging bounding boxes. The accuracy of the analysis can only be confirmed by the person supervising the analysis.

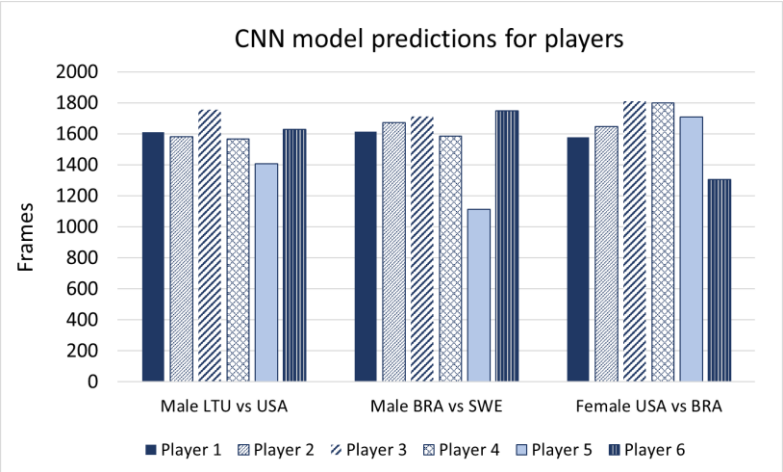

Figure 10: CNN based player tracking results with different video streams

5 Conclusions

In this paper, the research of object tracking techniques used for real-time goalball video analysis has been performed. Our proposal's novelty is that we have adapted single object tracking algorithms to solve multiple object tracking task. We also applied multiple object tracking models to the analysis of sports video material. We carried out different experiments to evaluate two multiple objects tracking task approaches: by employing a single object tracker (including Boosting, CSR-DCF, KCF, MOSSE, TLD); and CNN for multiple object tracking. For the first approach, we evaluate the method's performance in terms of the number of frames and speed. Experiments have shown that only the KCF algorithm can determine the adjustments of a player's position. MOSSE algorithm outperforms other algorithms in terms of speed and is three times faster than KCF and 9,8 times faster than TLD. CNN results are promising for players' position prediction, and accuracy varies from 88.12% to 90,26%; the accuracy was measured by calculating the total number of frames where each player was predicted and classified. However, CNN has shown poor performance for ball predictions providing 39% average accuracy of ball position. An interesting direction for further research would be to combine neural networks-based object detection and single object tracking in order to get better tracking results.

6 Acknowledgments

We want to express our very great appreciation to Dr. Agnė Paulauskaitė-Tarasevičienė for her insights and advice.

7 References

[1] E. Martinez-Martin ir A. P. d. Pobil, „Object Detection and Recognition for AssistiveRobots,“ *Robotics & automation magazine,* t. 24, pp. 123 - 138, 2017.

[2] M. S. Adam, M. H. Anisi ir IhsanAli, „Object tracking sensor networks in smart cities: Taxonomy, architecture, applications, research challenges and future directions,“ *Future Generation Computer Systems,* t. 107, pp. 909 - 923, 2020, June.

[3] F. Joy ir V. V. Kumar, „A review on multiple object detection and tracking in smart city video analytics,“ *Research gate,* 2018, January.

[4] M. Li, „Detecting, segmenting and tracking bio-medical objects,“ *Scolars Mine Doctoral Dissertations,* 2016.

[5] Y. Wang, B. Georgescu, T. Chen, W. Wu, P. Wang, X. Lu, R. Ionasec, Y. Zheng ir D. Comaniciu, „Learning-Based Detection and Tracking in Medical Imaging: A Probabilistic Approach,“ *M. González Hidalgo et al. (eds.), Deformation Models, Lecture Notes in Computational Vision and Biomechanics 7,* pp. 209 - 235, 2013.

[6] R. Azad, B. Azad, N. B. Khalifa ir S. Jamali, „Real-time human-computer interaction based on face and hand gesture recognition,“ *International Journal in Foundations of Computer Science & Technology (IJFCST),* t. 4, nr. 4, pp. 37 - 48, 2014, July.

[7] Z. Kalal, K. Mikolajczyk ir J. Matas, „Tracking-learning-detection. Pattern Analysis and Machine Intelligence,“ *IEEE Transactions,* pp. 1409 - 1422, 2012.

[8] „SAP and Panasonic Launch Joint Initiative for Video-Based Sports Analytics Solutions,“ *SAP News,* 2014, September 12.

[9] M. Danelljan, G. Hager, F. S. Khan ir M. Felsberg, „Accurate scale estimation for robust visual tracking,“ *roc. British Machine Vision Conference,* %1 t. iš %21, 2, 4, 8, pp. 1 - 11, 2014.

[10] Danelljan, G. Hager, F. S. Khan ir M. Felsberg, „Learning spatially regularized correlation filters for visual tracking. Pages 4310 - 4318,“ įtraukta *IEEE International Conference on Computer Vision,* Santiago, Chile, 2015, December 7-13.

[11] M. Danelljan, F. S. Khan, M. Felsberg ir J. v. d. Weijer, „Adaptive color attributes for real-time visual tracking. Pages 1090–1097,“ įtraukta *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, June 23 - 28.

[12] Z. Hong, Z. Chen, C. Wang, X. Mei, D. Prokhorov ir D. Tao, „Multi-store tracker (muster): A cognitive psychology inspired approach to object tracking,“ *Comp. Vis. Patt. Recognition,* pp. 749 - 758, June 2015.

[13] H. Grabner, M. Grabner ir H. Bischof, „Real-time tracking via on-line boosting,“ *BMVC,* t. 1, p. 6, 2006.

[14] A. Lukezic, T. Voj'ir, L. C. Zajc, J. Matas ir M. Kristan, „Discriminative correlation filter tracker with channel and spatial reliability,“ *International Journal of Computer Vision,* 2018.

[15] J. F. Henriques, R. Caseiro, P. Martins ir J. Batista, „Exploiting the circulant structure of tracking-by-detection with kernels,“ *In proceedings of the European Conference on Computer Vision,* 2012.

[16] D. S. Bolme, J. R. Beveridge, B. A. Draper ir M. L. Yui, „Visual object tracking using adaptive correlation filters,“ įtraukta *Computer Vision and Pattern Recognition (CVPR),* 2010.

[17] S. K. S. Anjali B Guptha, „Multiple Face Detection and Tracking using Viola-Jones Algorithm,“ *International Research Journal of Engineering and Technology (IRJET),* t. 07, nr. 04, 2020.

[18] R. Padilla, S. L. Netto ir E. A. B. d. Silva, „A Survey on Performance Metrics for Object-Detection Algorithms,“ 2020.