Recent Developments in Tracking Objects in a Video Sequence

Michał Staniszewski[1,2], Mateusz Kloszczyk[1], Jakub Segen[1],
Kamil Wereszczyński[1,2], Aldona Drabik[1], and Marek Kulbacki[1](✉)

[1] Polish-Japanese Academy of Information Technology,
Koszykowa 86, 02-008 Warszawa, Poland
mk@pjwstk.edu.pl

[2] Institute of Informatics, Silesian University of Technology,
Akademicka 16, 44-100 Gliwice, Poland

Abstract. Methods of tracking of multiple objects or people in video sequences have applications in many fields such as surveillance, art, transport or biology. This, over four decades old area is still very active, with multiple new contributions presented every year. Tracking methods must solve intricate problems, for example occlusion of many objects, crowded scenes, illumination of different places and motion of camera. This paper presents a brief survey of recent developments in video tracking based methods, focused mainly on the last three years. The surveyed methods are divided into two groups: tracking by detection, which includes methods that solve the problem of time-linking objects detected in all video frames, and tracking by correlation, containing methods that follow a selected object using cross correlation. The reviewed methods are collected in a table that lists for each method the benchmark datasets used for its evaluation, implementation environment, and whether it can track single or multiple objects.

Keywords: Object tracking · Computer vision · Statistical analysis · Video signal processing

1 Introduction

In the last four decades many approaches were designed in the field of tracking of multiple objects, which can be applied in surveillance, detecting human behavior and many other aspects of computer vision. There are many problems in this area that should be taken into consideration. In real tracking problems people may be occluded by other persons or objects. Video sequences may be influenced by different kind of light and illumination which lead to many changes in colors. Finally camera can move and the same person or object should be still traced. The development of tracking approaches from last decade led to creation of many methods having different type of input and output. Current methods were tested also on different data sets which is not reliable for larger evaluation. That problem was introduced in [3] and as a result authors collected and shared

N.T. Nguyen et al. (Eds.): ACIIDS 2016, Part II, LNAI 9622, pp. 427–436, 2016.
DOI: 10.1007/978-3-662-49390-8_42

publicly available data benchmarks in one place with important annotations consisting of many different factors affecting tracking performance.

The tracking algorithms may be divided into few parts, which can be formulated in many different ways. Representation model of object can be described in the manner of sparse representation presented and improved in [11]. Additionally tracking algorithms may use visual features such as histograms of oriented gradients [5] or color [4] and Haar-like features [6]. In order to discriminate the target many learning methods have been applied such as Support Vector Machines (SVM) with its modifications [7] or boosting methods [8]. In order to find the target localization deterministic [4] and stochastic [8] methods have been used. To deal with appearance changes the object model has to be updated which can be done effectively by proposed algorithms of online mixture model [9] or online boosting [6]. Surrounding context is also important in discriminating objects in occlusion which has been presented in [10] (Fig. 1).

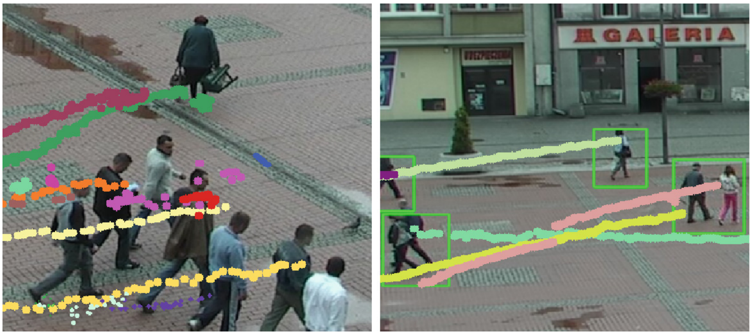

Fig. 1. Possible problems that may be present during procedure of tracking. On the left side objects traveling in one group, on the right side many crossing tracks with possibility of occlusions.

Contributions of that paper into tracking problem rely on comparison of methods from last three years which have shared source code on project's web page (that gives opportunity to real comparison). Authors presented mainly methods that solve the most challenging problem which is tracking of multiple objects. The paper consists also of methods dealing only with single object but using correlation filters. In comparison to other current surveys that paper is much more comprehensive [1,2] because it describes public benchmarks and much more methods and parameters of evaluation.

The paper is organized in following way - in the second section authors describe current state-of-the-art in the groups of tracking by detection and by use of correlation filters. In the third and fourth sections authors present available data benchmarks containing video sequence and possible parameters used in order to mark investigated methods.

2 Current Methods

The section is built from descriptions of few methods which were divided into two groups. The first one contains methods based on tracking by detection while the second one methods using correlation filters. All of them have available source code which can be downloaded and tested. Authors added also few comments for each method in the order of advantages and disadvantages.

2.1 Tracking by Detection

Group of methods located in the term of tracking by detection has one common feature. The possible approach contains of application of discriminative appearance learning. Such methods contain online learning by use of appearance prediction of given object in the next frame of video. Such prediction can be obtained thanks to algorithms of SVM [7], random forest methods or boosting models [8]. A number of these methods link detections in different frames by solving an optimization problem defined on a graph. It can be solved either locally by application of hungarian algorithm where the problem is approximated or globally where instead of few frames, methods use batch of frames getting more global overview. Additionally methods of tracking by detection can have two categories - batch and online. Batch methods use in the analysis detections of all frames within batch and connect short parts (tracklets) in longer parts (tracks). Such approach requires much bigger computational power. On the other hand online methods can be applied for many issues in the real time by sequencing building of trajectories in connection frame by frame.

Tracking Multiple People Online and in Real Time [12]**.** The method of Online tracking of multiple objects uses a multi-stage cascade combined with a sliding temporal window. In fact input of given source code consists of videos and set of detections which can be described by its appearance feature, position in time and estimated velocity that object moves. In order to describe appearance of a person in algorithm an HSV color histogram is applied, but algorithm can be extended with many other descriptors without any modification of that method. For each pair of observations (coming from person detection) algorithm measures the evidence of being or not identical. The value of evidence is computed basing on data and calculated as a measure of correlation, which can be a number coming from the set $(-\text{inf}, [-1, 1], +\text{inf})$. If the pair of observations evidence is indicated by positive value while negative shows evidence against and zero is connected to indifference. Infinite (negative and positive) corresponds to hard constraints. The scheme of algorithm consists of two simple phases built from two stages each. In the first part partial detections are connected to form of short tracklets (in short time) basing on appearance and space-time affinities. In the second part method operates on entire temporal window which results in whole trajectories. **Advantages** Approach similar to GMCP [13], multi person tracking is done jointly for all identities (in contrast to sequential GMCP), result depends mainly on number of observations (not on length of sequence), implementation consists of separate parts connected to detections and tracklets. **Disadvantages** Formulation of correlation matrix used as a distance and similarity comperator causes time complexity.

The Way they Move: Tracking Multiple Targets with Similar Appearance [14]**.** The algorithm is computationally efficient and is dedicated for tracking of multi-object by detection that can overcome four main problems: camera movement, appearance similarity of many targets, lack of data due to being out of the field or occlusion, crossing trajectories. The method takes as input a set of short tracklets that have different lengths and no appearance information. The main problem is related with connection of tracklets that belong to similar trajectories. In this paper authors propose set of methods using dynamics of motion that can handle many objects described by long tracks with possibility of occlusion and lack of data. In the first part the problem is formulated as a generalized linear assignment (GLA) [15] of short tracklets which are used to built longer trajectories by use of similarity of motion which needs efficient algorithms in order to estimate such similarities. The algorithm does not require any prior assumption connected to starting and ending nodes or length of trajectories. Estimation of dynamics similarity can be done thanks to two algorithms proposed by authors. The first one, presented in paper, alternating direction method of multipliers (ADMM) [16] uses different optimization which has lower memory requirements and relies on solving a relaxation and assumes choice of parameters responsible for noise penalty and estimation of the rank basing on singular values. In second case authors propose new algorithm basing on iterative Hankel Total Least Squares (IHTLS) [17] which is connected in finding the rank of noisy data formulated in Hankel matrices. That algorithm was contributed by authors to clean noisy matrices and estimate rank. **Advantages** Method allows tracks to start and terminate anywhere in position and time, algorithm can operate at the tracklet level, but also all the data on a tracklet rather than a selected portion, does not assume any priors for the target motion. **Disadvantages** Algorithm should be compared with state-of-the-art tracking method basing on currently available benchmarks in order to check its reliability.

GMMCP Tracker: Globally Optimal Generalized Maximum Multi Clique Problem for Multiple Object Tracking [18]**.** The method of tracking is formulated as a data association of Generalized Maximum Multi Clique problem (GMMCP). Input for the algorithm is based on finding detection in each frame described in [31] to get the detection hypothesis in each frame. Long tracks are built from shorter tracklets in two layer framework. GMMCP starts with low level tracklets which are limited to a maximum of 10 frames due to elimination of tracklets shorter than 5 frames. In the next step low level tracklets in segments create the input which will form the first layer of framework returning mid level tracklets, which are used later in forming the final trajectories. In comparison to different tracklets, algorithm takes into consideration appearance affinity, using for each node color histogram. The histogram is calculated for each detection and the representation of the nodes are computed by the median appearance. In order to find the appearance affinity of two tracklets algorithm calculates the histogram intersection. Additionally in tracking method motion model is taken in the form of constant velocity. **Advantages** Contribution to GMCP [13] method as a new graph theoretic problem, efficient occlusion handling, improvement in comparison to other methods, performance close to real time. **Disadvantages** The time complexity increases with the number of objects.

Joint Tracking and Segmentation of Multiple Targets [19]**.** The algorithm is dedicated for the problem of multi-target tracking that can exploit low level image information and connect each pixel to an investigated object or treats it as background. The result of the method in the form of video segmentation can work in real world videos. The algorithm takes as an input video sequence and starts with assigning a unique ID not only for each target detection but also to specially defined superpixel in videos. Thanks to that idea the method can be used in recovery of trajectories in long occlusion due to existence of superpixels even when the detections are missed. Authors used multi-label conditional random field (CRF) in order to model the problem, which is stated as finding hypotheses of long trajectories that best describe the data basing on the low-level information. **Advantages** Improvement by 0,1 on average of missing recalls, while reduction of the number of ID switches. **Disadvantages** Needs improvement in order to work in time close to real time.

Tracking Multiple High-Density Homogeneous Targets [20]**.** The method was created to deal with detection and tracking multiple objects which are dense and homogeneous. Objects detection is made by a technique of gradients and usage of isocontour for intensity maps. The method of tracking recursively connects detection by use of graph in temporal window, which is solved by greedy algorithm. Problem of detection was solved by automatic finding of objects thanks to local maximum. Additionally detector uses isocontour on intensity maps, which contributes connection with objects. Tracking works on short tracklets and by use of greedy algorithm validates results backward in temporal window. Authors added possibility of online reduction of false tracks in connection step. The whole process consists of connection of detection in time, forecasting detection in particular frames and reduction of false detections. At the beginning short tracks are generated by Hungarian algorithm from those detections having the biggest affinities. Long trajectories are built from short tracks with probability of connection. **Advantages** Detection is background independent and does not require learning of model features such as color and texture. Good results for detections and tracking for prepared datasets. **Disadvantages** Complex algorithm with good results however not so competetive. For public benchmarks observable changes in ID switch.

Robust Online Multi-object Tracking Based on Tracklet Confidence and Online Discriminative Appearance Learning [21]**.** The main goal of the method was to create algorithm that will track multiple objects basing on tracklet confidence, which is evaluated on the base of detection and continuity. In the next step detections are connected into tracklets by use of Online discriminative appearance learning getting difference in objects. In order to face with occlusion across whole video sequence method uses tracklet confidence and the whole problem is solved depending on value of confidence - reliable tracklets with high confidence are connected locally while low tracklets with low confidence come back into remaining and not finished tracklets and detections. The crucial moment in presented method relies on connecting tracklets globally and locally in order to form one object. For that reason author proposed application of algorithm of Online discriminative appearance learning which is used for updating model according to results of tracking and online training of collecting data in order to distinguish occurring new objects. Such assumption can be satisfied by Incremental Linear Discriminant Analysis method (ILDA) [22], which helps in distinguishing many objects and updating model. **Advantages** Algorithm simple, intuitive and well tested. Application of tracklet confidence and discriminative online learning. **Disadvantages** Matlab implementation needs to be optimized, large time complexity.

2.2 Tracking by Correlation

In last years tracking has been enriched by methods of correlation filters. In comparison to tracking by detection methods, the results of application of correlation filters lead to less computational load even on hundreds of frames. The main advantage of that group lies in the lack of necessity in iterating over all objects and also in application of Fourier domain.

Accurate Scale Estimation for Robust Visual Tracking [23]**.** In contrast to other method that algorithm strongly takes into consideration robust scale estimation for visual tracking. The method start with videos formulated as raw pixels or Histograms of Oriented Gradients (HOG) [28] and uses learning discriminative correlation filters working on a representation of scale pyramid. The improvement of scale searching bases on learning different filters for translation and scale estimation. The method contributes the discriminative correlation filters first time applied in the MOSSE tracker [24]. One of the advantages relies on incorporating idea of scale estimation on other tracking frameworks. **Advantages** Method is more than 2.5 times faster than Struck [25], 25 times faster than Adaptive Structural Local Sparse Appearance (ASLA) [26] and 250 times faster than Sparsity Collaborative Model (SCM) [27] in median frame per second (FPS). **Disadvantages** Results are presented for single moving object and needs modification in order to serve for multiple objects.

Fast Visual Tracking via Dense Spatio-Temporal Context Learning [29]**.** The presented robust algorithm uses the dense spatio-temporal context for visual tracking and takes video sequence as an input. In the first part of algorithm a model of spatial context is evaluated between the object and its local background by means of spatial correlations by solving deconvolution problem. In the next frame the learned spatial context is used in order to update the model. Building trajectories is made by formulating a confidence map as a convolution problem. The best possible object position can be computed by getting maximum of the confidence map and adapting a novel scale estimation scheme which results in final track. **Advantages** The proposed algorithm is simple and fast that needs only 4 computation of Fast Fourier Transforms (FFT) at 350 FPS in MATLAB. Method uses explicit scale update **Disadvantages** Method was tested mainly on data prepared by authors.

High-Speed Tracking with Kernelized Correlation Filters [30]**.** The method uses a correlation filter in tracking problem for videos built from raw pixels or HOG descriptors [28]. The presented tool can operate on thousands of objects with different relative translations without a need of iterating on them. That solution works in the Fourier domain where some learning algorithms can operate even if new samples are added for specific model. Correlation filters use a feature that the convolution of two samples (dot-product at different translations) is equal to an element-wise product performed in the Fourier domain. Thanks to that advantage and application of the Fourier domain, the desired results of linear classifier can overcome problems for many translations and image shifts. **Advantages** Taking advantage of all 4 cores of a desktop computer, method take less than 2 min to process all 50 videos (29,000 frames). **Disadvantages** Lack of comparison to available benchmarks (Table 1).

Table 1. Summary of all described methods with respect to benchmarks applied for evaluation and statement of used code. Additionally possibility of method to deal either with single object or multi objects.

Method	Benchmarks	Code	Object
Tracking multiple people online in real time [12]	Towncenter [33], pets2009 [32], Parkinglot [13]	Matlab	Multi object
The way they move: tracking multiple targets [14]	Own data [14], TUD [34]	Matlab	Multi object
GMMCP tracker [18]	TUD [34], Parkinglot 2 [18]	Matlab	Multi object
Joint tracking and segmentation of multiple targets [19]	TUD campus [34]	Matlab	Multi object
Tracking multiple high-density homogeneous targets [20]	ETH [35], pets [32], TUD [34]	Matlab	Multi object
Multi-object tracking based on tracklet confidence [21]	ETH [35]	Matlab	Multi object
Accurate scale estimation for robust visual tracking [23]	Own data [23]	Matlab	Single object
Fast visual tracking via dense spatio context learning [29]	Own data [29]	Matlab	Single object
High-speed tracking with kernelized correlation filters [30]	Own data [30]	Matlab	Single object

3 Available Data Benchmarks

There are many currently used video benchmarks that are tested in order to compare presented methods. Thanks to such dataset already published methods can be compared in terms of similar aspects. All of benchmarks are built from many videos or set of pictures that are presenting different scenes with partial occlusion and other tracking problem. Some of them consist of ground truth table and detections that can be used as a reference. The most popular videos are stored in benchmarks:pets2009 [32], towncenter [33], Parkinglot and Parkinglot 2 [13,18], TUD (Crossing, Campus, Stadtmitte) [34], ETH (Bahnhof, Jelmoli, Sunny Day) [35], videos from own dataset and youtube containg moving objects, sport scenes etc. [14].

4 Methods of Evaluation

Authors of previous works used many different parameters which are used for comparison purposes. Most of them are based on a specially prepared ground truth table which consists of hand made detections. The most popular parameters are listed below.

- increase the number of false positive detections by adding random detections into the set or increase the number of false negative detections by removing correct detections from the set [14],
- number of ID switches when trajectories are crossing [14],
- Score as a relation of area of common part of detected bounding box and ground truth to their sum [14],
- distance precision calculated as the average Euclidean distance between the estimated centre location of the target and the ground-truth [23],
- centre location error computed as the relative number of frames in the sequence where the location error is smaller than a certain threshold [23],
- overlap precision defined as the percentage of frames where the bounding box overlap surpasses a threshold [23],
- multiple Object Tracking Accuracy (MOTA) combines the number of false positives, false negatives and identity switches over all frame indices [12].

5 Conclusion

New methods of tracking multiple objects are being developed at an even pace and the field remains competitive. The present survey of recent work in the area shows that current methods can be very effective in a wide range of environments and imaging conditions. However, there is still much room for improvement, especially in cases of multiple occlusions and where there are many objects in close proximity. Also, the speed of execution needed for real time tracking remains a challenge for the computationally more demanding methods. The continuing work will be devoted to a precise performance comparison among the current methods, by testing them on a unified benchmarking dataset and using the same configuration parameters.

Acknowledgments. This work has been supported by the National Centre for Research and Development (project UOD-DEM-1-183/001 “Intelligent video analysis system for behavior and event recognition in surveillance networks”).

References

1. Li, X., Hu, W., Shen, C., Zhang, Z., Dick, A., van den Hengel, A.: A Survey of Appearance Models in Visual Object Tracking (2013). CoRR abs/1303.4803
2. Chu, D.M., Cucchiara, R., Calderara, S., Dehghan, A., Shah, M.: Visual tracking: an experimental survey. Pat. An. Mach. Intel. **36**, 1442–1468 (2013)
3. Wu, Y., Lim, J., Yang, M.-H.: Online Object Tracking: A Benchmark CVpPR 2013, pp. 2411–2418 (2013). http://visual-tracking.net
4. Comaniciu, D., Ramesh, V., Meer, P.: Kernel-based object tracking. PAMI **25**(5), 564–577 (2003)
5. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
6. Grabner, H., Grabner, M., Bischof, H.: Real-time tracking via on-line boosting. In: BMVC (2006)
7. Avidan, S.: Support vector tracking. PAMI **26**(8), 1064–1072 (2004)
8. Babenko, B., Yang, M.-H., Belongie, S.: Visual tracking with online multiple instance learning. In: CVPR (2009)
9. Jepson, A.D., Fleet, D.J., El-Maraghi, T.F.: Robust online appearance models for visual tracking. PAMI **25**(10), 1296–1311 (2003)
10. Santner, J., Leistner, C., Saffari, A., Pock, T., Bischof, H.: PROST: parallel robust online simple tracking. In: CVPR (2010)
11. Mei, X., Ling, H.: Robust visual tracking using L1 minimization. In: ICCV (2009)
12. Ristani, E., Tomasi, C.: Tracking multiple people online and in real time. In: 12th Asian Conference on Computer Vision, pp. 444–459 (2014). https://www.cs.duke.edu/ristani/bip_tracker.html
13. Zamir, A.R., Dehghan, A., Shah, M.: GMCP-Tracker: global multi-object tracking using generalized minimum clique graphs. In: Proceedings of the 12th European Conference on Computer Vision, pp. 343–356 (2012). http://crcv.ucf.edu/projects/GMCP-Tracker/
14. Dicle, C., Camps, O., Sznaier, M.: The Way They Move: Tracking Multiple Targets with Similar Appearance Computer Vision (ICCV) (2013). https://bitbucket.org/cdicle/smot
15. Rossand, G., Soland, R.: A branch and bound algorithm for the generalized assignment problem. Math. Program. **8**(1), 91–103 (1975)
16. Ayazoglu, M., Sznaier, M., Camps, O.: Fast algorithms for structured robust principal component analysis. In: CVPR, pp. 1704–1711 (2012)
17. Park, H., Zhang, L., Rosen, J.: Low rank approximation of a hankel matrix by structured total least norm. BIT Numer. Math. **39**(4), 757–779 (1999)
18. Dehghan, A., Assari, S., Shah, M.: GMMCP tracker: globally optimal generalized maximum multi clique problem for multiple object tracking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4091–4099 (2015). http://crcv.ucf.edu/projects/GMMCP-Tracker/
19. Milan, A., Leal-Taixe, L., Schindler, K., Reid, I.: Joint tracking and segmentation of multiple targets CVPR (2015). https://bitbucket.org/amilan/segtracking
20. Poiesi, F., Cavallaro, A.: Tracking multiple high-density homogeneous targets. IEEE Trans. Circ. Syst. Video Technol. **25**, 623–637 (2015). http://www.eecs.qmul.ac.uk/andrea/thdt.html
21. Bae, S.-H., Yoon, K.-J.: Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning CVPR, pp. 1218–1225 (2014). https://cvl.gist.ac.kr/project/cmot.html

22. Kim, T.-K., Stenger, B., Kittler, J., Cipolla, R.: Incremental linear discriminant analysis using sufficient spanning sets and its applications. IJCV **91**(2), 216–232 (2011)
23. Danelljan, M., Hager, G., Shahbaz, K., F., Felsberg, M.: Accurate scale estimation for robust visual tracking. In: Proceedings of the British Machine Vision Conference (2014). http://www.cvl.isy.liu.se/en/research/objrec/visualtracking/scalvistrack/index.html
24. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: Computer Vision and Pattern Recognition (2010)
25. Hare, S., Saffari, A., Torr, P.: Struck: structured output tracking with kernels. In: Computer Vision and Pattern Recognition (2011)
26. Jia, X., Lu, H., Yang, M.-H.: Visual tracking via adaptive structural local sparse appearance model. In: Computer Vision and Pattern Recognition (2012)
27. Zhong, W., Lu, H., Yang, M.-H.: Robust object tracking via sparsity based collaborative model. In: Computer Vision and Pattern Recognition (2012)
28. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition, pp. 886–893 (2005)
29. Zhang, K., Zhang, L., Liu, Q., Zhang, D., Yang, M.H.: Fast visual tracking via dense spatio-temporal context learning. In: 13th European Conference, Zurich, pp. 127–141 (2014). http://www4.comp.polyu.edu.hk/cslzhang/STC/STC.htm
30. Henriques, J.F., Caseiro, R., Martins, P., Batista J.: High-Speed Tracking with Kernelized Correlation Filters, CoRR (2014). abs/1404.7584http://home.isr.uc.pt/henriques/circulant/
31. Felzenszwalb, P., Girshick, R., McAllester, B., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intell. **32**(9), 1627–1645 (2010)
32. Ferryman, J.: Proceedings (pets 2009). Eleventh IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (2009)
33. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: Computer Vision and Pattern Recognition (2011)
34. Andriluka, M., Roth, S., Schiele, B.: People-tracking-bydetectionandpeople-detection-by-tracking. In: Proceedings of IEEE Conference Computer Vision and Pattern Recognition, pp. 1–8 (2008)
35. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: Computer Vision and Pattern Recognition (2008)