# A Robust Online Multi-Face Tracking System

Martin Soldić, Darijan Marčetić, Slobodan Ribarić
University of Zagreb, Faculty of Electrical Engineering and Computing
{martin.soldic, darijan.marcetic, slobodan.ribaric}@fer.hr

***Abstract*— This paper presents a system for robust online multi-face tracking in video sequences recorded by a stationary camera. Several components and algorithms are combined in the system architecture: an NPD robust face detector, a DSST tracker augmented with FIFO long- and short-term memories (LTMs/STMs), trajectory memories (TMs), a PSR-based tracking failure detector and a Hungarian algorithm for trajectory assignment. The paper gives the preliminary experimental results, expressed by testing MOT, MOTP and IDS metrics, obtained on the benchmark dataset consisting of three videos.**

***Keywords—face tracking, NPD detector, DSST tracking, PSR metric, online tracking***

## I. INTRODUCTION

Multiple target tracking (MTT) estimates locations and scales of multiple targets in every frame of a video and preserves a label of each target. The common applications of MTT are the tracking of pedestrians [1], sport players, vehicles on the road and animals [2, 3, 4]. Pedestrian tracking is still the most frequent objective [1] of MTT. MTT algorithms are far below human performance when objects have a complex appearance, move at high speed, have large occlusions and/or interact with objects in the scene. The best results in MTT are achieved when the tracking by detection (TBD) paradigm [5, 6, 7, 8] is used, where a specialized object detector is used in every frame of a video and all possible trajectories are optimized to find the best object trajectories. TBD can be classified into two main categories: online and offline tracking methods. Online tracking estimates the states (locations, scales) of targets in each frame by inferring from past and current observations. Online tracking methods can be further divided into probabilistic and determinist categories. The most common probabilistic methods are: the particle filter [9], the JPDA (*Joint Probabilistic Data Association*) [10] and the single-scan MCMCDA (*Monte Carlo Markov Chain Data Association*) [11], while the deterministic methods are greedy assignment and bipartite matching with the Hungarian algorithm [12]. The main advantage of these methods is a short time delay when optimizing trajectories while the main drawback is that the trajectories cannot be corrected once they are generated.

Offline methods, attempt to find the best possible trajectory assignments for the given optimization technique taking into consideration all the sliding windows of all frames in a video. The time delay for offline methods is equal to the total time needed to generate trajectories for the whole video. Their application is thus limited to the video analysis of past events and they cannot be used for live tracking. Popular offline methods are based on graph flow optimization [7], linear programming [6], multi-clique [8], multi-scan MCMCDA [11] and discrete/continuous energy optimization Milan *et al.* [13].

Another important category for multi target tracking is detection free based methods [14, 15], which do not involve object detection during tracking. The only requirement is to select, manually or by a detector, targets in the first frame. These methods assume a fixed number of targets during the entire video sequence and cannot track a new target when it enters a scene. The advantage of these methods when compared to a TBD is that they do not use a detector (which is time consuming) and the results of the tracking do not depend on the performance of a pre-trained detector.

This paper focuses on multi-face tracking in video recorded by one stationary camera with no restrictions on background complexity, where it is assumed that new faces after initialization cannot appear during each video sequence. The robustness of the MTT tracking is mainly related to minimizing the number of failure tracking and identity (i.e. label) switches during the abrupt appearance or motion changes of the tracked objects, in the cases of long- and/or short-term full occlusions. The proposed MTT is online, i.e. without trajectory optimization based on previous frames (trajectories are generated on the fly). The proposed MTT is a component of the face de-identification pipeline where there are requirements for robust face detection in each frame. The combination of face detection and tracking, i.e., the combination of the spatial and temporal correspondence between frames, can improve the effectiveness of the detection and localization of faces, and can also remove false positive detections.

Trajectory assignments based on a combination of re-detected faces and the Hungarian algorithm are used to recover tracklet(s) in the case of tracking failure. For initialization and redetection, the fast and accurate unconstrained face detector NPD (*Normalized Pixel Differences*) [16] is used. Multiple instances of the DSST (*Discriminative Scale and Space Tracking*) [17] tracker are used for multi-face tracking. The rest of the paper is organised as follows: section 2 presents the main components of the tracking system, section 3 presents the robust multi-face tracking system in detail and the experimental setup and results are given in section 4.

## II. BACKGROUND

The architecture is composed of the following main components: the NPD face detector, multiple DSST tracers, the target lost detector based on PSR (Peak-to Sidelobe Ratio) [18], and metric and trajectory assignment based on the Hungarian algorithm [12].

### A. Face detector

A NPD (*Normalized Pixel Differences*) detector [16] is a robust and fast face detector with high recall, based on a simple feature and cascade of deep quadratic trees, where the

AdaBoost algorithm is used to select the most discriminative features. The NPD features are defined as: $f(v_{i,j}, v_{k,l}) = (v_{i,j} - v_{k,l}) / (v_{i,j} + v_{k,l})$, where $v_{i,j}$ , $v_{k,l} \geq 0$ are intensity values of two pixels at the positions $(i, j)$ and $(k, l)$. A quadratic splitting strategy for deep tree, where the depth is eight is used for learning. The values of the NPD features are quantized into L = 256 bins. The strong classifier is based on 1226 quadratic trees and 46401 NPD features. There are only 114 average evaluations of features per detection window. The output of the NPD are squares representing face candidates with values of confidence.

### B. *Tracking algorithm*

DSST (*Discriminative Scale and Space*) tracker [17] finds an optimal target position and target scale using discriminative position and scale correlation filters, respectively. An exhaustive search in position-scale space is avoided by applying a position filter first and then a scale filter on an area centred on position filter maximum score. Both filters are based on F-HOG [19] features. The circular correlation in the Fourier domain is used for computing scores between target features in the current frame and learned filters. During learning, the position and scale filters from the current frame contribute by factor η, while the filters from previous frames contribute by factor (1- η). The value of eta is 0.025. Both position and scale filters are separately calculated. This learning procedure makes the algorithm more robust to target appearance change. The algorithm achieved state-of-the-art performance on the VOT2014 benchmark [20]. The periodic assumption in the circular correlation introduces negative boundary effects on patches used in training, thus reducing the discriminative tracking efficiency. This dramatically reduces performance if the target moves fast and/or undergoes long-term occlusion.

### C. *Tracking failure detection*

The PSR (Peak-to Sidelobe ratio) [18] metric is used for tracking failure detection. The response map obtained from the position correlation filter is divided into a square area around the peak and the remaining area, called the sidelobe. The square area around the peak is approximately 12 % of the size of the response map.

The PSR metric is calculated from the following:

$$PSR = \frac{(MaxPeak - \mu)}{\sigma} \qquad (2.1.)$$

where mean (μ) and standard deviation (σ) are calculated from the values in the sidelobe area. Typical good tracking PSR values range from 20 to 60 and tracking failure is considered if PSR < 10.

### D. *Assignment method*

The Hungarian assignment [12] algorithm is a simple assignment strategy that minimizes the cost of assigning detections to trajectories by only allowing a one to one assignment. It is based on dynamic programming and finds its optimum if the measurements are reliable.

## III. ROBUST MULTI-FACE TRACKING SYSTEM

The architecture of the robust multi-face tracking system is depicted in Figure 1. The proposed tracking system is composed of five main stages: 1) Initialization and faces detection by the NPD detector (Section A); 2) Multi-face tracking with multiple DSST trackers (Section B); 3) Tracking failure detection based on the PSR metric (Section C); 4) Proposed tracking procedure (Section D); 5) Assignment procedure (Section E).

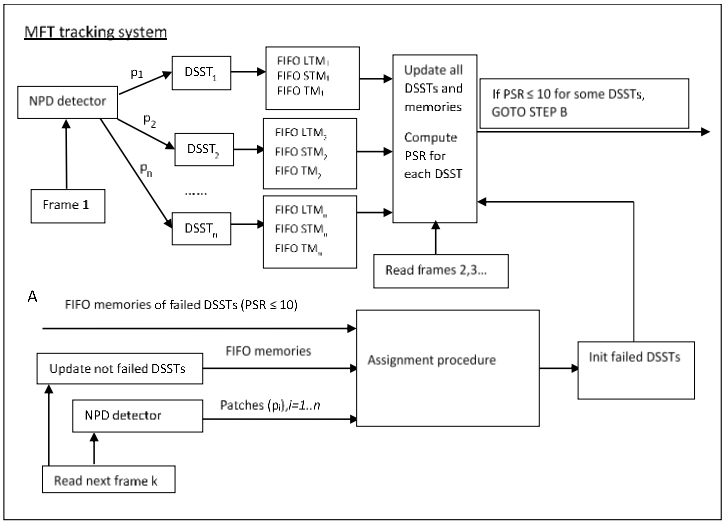


Fig. 1. Architecture of the multi-face tracking system.

### A. *Face detection*

For automatic multiple DSST trackers initialization, the NPD detector is used in the first frame of the video (or in consecutive frames until some face candidates i.e. patches, are found). Each patch $p_i$; $i$ = 1, 2, ..., $n$ returned by the NPD is appropriately extended (10 % of patch size) in all directions to guarantee capturing the entire face. The NPD is also used for re-initialization when some of the DSST trackers fails.

### B. *DSST tracker with memories*

Every DSST tracker, based on the F-HOG scale and position filters, tracks a face until abrupt face motion and short and/or long-term occlusions occur. During tracking, the current position filter is stored in a FIFO STM, while an aggregate position filter is periodically stored (e.g. every t units) in FIFO LTM.

Every DSST tracker has an additional trajectory memory TM, which is used for storing the trajectory represented by the coordinates of the patch centre, i.e. *tracking window*.

Tracking failure can be caused by short- or long-term full-face occlusion, incorrect target position/scale estimation due to an abrupt appearance and motion change or when the tracked face leaves a scene. Recovery from a tracking failure is supported by the contents of memories (STM, LTM and TM). LTM and STM organized as FIFO have depths $d_{LTM}$ and $d_{STM}$, respectively. The LTM stores an aggregated position correlation filter every $c_{LTM} > 1$ frames, while the STM stores the current position correlation filter calculated each time

during tracking. Due to the discriminatory weakness of the F-HOG features, the position correlation filters cannot be used to explicitly distinguish faces based on recognition [21], thus making them unreliable for the explicit tracking recovery of several faces when they are occluded and lost. The above problem is solved in the following manner: The position correlation filters stored in the STM of DSST trackers that have not failed are used for matching with the F-HOG features of NPD re-detected patches. The combination of the result of matching and the intersection over union (IOU) criterion is used to assign a corresponding patch (face) to a corresponding tracker. Additionally, to reduce false positives in image patches returned by the NPD detector at the re-detection stage, the position correlation filters stored in the LTM are used [22]. Then, the FIFO TM with $d_{TM}$ depth is used for storing the centre of the tracking window in each frame. These position data are used for position estimation in assignment after the target gets lost (see Section 3.E).
Memory structures are depicted in Fig. 2.

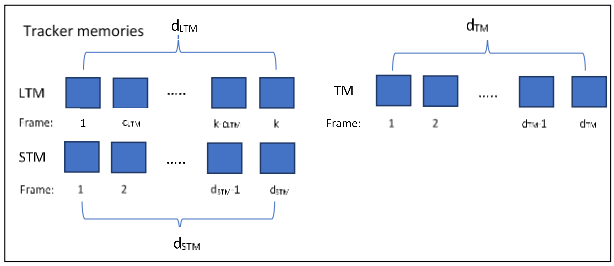


Fig. 2. FIFO memories used by the tracking system.

### C. Tracking failure detection

Tracking failure is determined by using the Peak to Side Lobe Ratio (PSR) metric [18] on a filter response obtained from the correlation between the aggregated position filter and the filter calculated on the current image during the tracking update stage. The details are described in section 2.C.

### D. The tracking procedure

The tracking system is based on the following algorithm:

**STEP 1**: The NPD face detector is invoked in the first frame and in all consecutive frames until at least one face is found. The face candidates obtained by the detector are used for initializing multiple DSST trackers (Section 3. A).

**STEP 2:** For each DSST tracker do the following: Initialize the FIFO LTM and STM for storing the position correlation filters. Initialize TM for the storing position coordinates (x,y) of the face.

**STEP 3**: Read a new frame. Determine new positions and scales of faces using DSST trackers. If some of the DSST trackers fail (PSR < 10), GO TO STEP 4, ELSE update all STMs with new position correlation filters, the LTMs (every $c_{LTM}$) with aggregated position correlation filters, and the TMs with new position values. GOTO STEP 3.

**STEP 4:** Re-detection procedure: Call the NPD detector on the current frame. IF face candidates are found THEN GOTO STEP 1 of the assignment procedure (Section E.). IF no face candidates are found THEN read a new frame and GOTO STEP 4.

**STEP 5:** Initialize the DSST trackers with face patches obtained in STEP 4 which should correspond to these trackers according to the assignment rules described in Section E. GOTO STEP 3.

### E. Assignment procedure

**STEP 1:** Do overlap matching between face patches returned by the NPD and the DSST trackers that have not failed. Use the intersection over union (IoU) criterion with threshold θ between each pair of NPD patches and the DSST tracker that has not failed. Pairs that have IOU > θ need no further analysis. For the rest of the pairs, do a correlation between the filters in the STM of the failed DSST trackers and image patches. Pairs with the highest response are assigned to each other and are not further considered. GOTO STEP 2 with only the failed DSST trackers and the remaining patches.

**STEP 2:** Do a correlation between the stored filter in the LTM of the failed DSST and the filters computed on the image patches. Remove all the candidates with PSR < 10. Patches that correspond to those that have been removed are false positives. GOTO STEP 3.

**STEP 3:** Do a position prediction based on linear regression by taking the trajectory points (previously stored in TM) for each failed DSST. GOTO STEP 4.

**STEP 4:** Calculate the distances between the predicted position of each DSST from STEP 3 and the centres of the image patches. Store these distances in a distance table. Do an optimal assignment using the Hungarian algorithm.GOTO STEP 5 of the tracking procedure.

## IV. EXPERIMENTAL RESULTS

The preliminary experimental results, expressed by the testing metrics [23]: MOTA (*Multiple Object Tracking Accuracy*), MOTP (*Multiple Object Tracking Precision*) and IDS (*Identity Switch*), were obtained on the benchmark dataset consisting of three videos: MotinasMultifront, MotinasMultiwild and Motinasfast [24].

These videos were recorded using a stationary camera, while the persons in videos moved around keeping their faces always in a frontal position - MotinasMultifront video (1269 frames) or in unconstrained positions - MotinasMultiwild (1007 frames) and Motinasfast (487 frames). Samples of frames are given in Fig. 3.

Fig. 3. Sample frames from each video: Multifront, Multiwild and Motinasfast, respectively. All three videos have 720 x 576 resolution.

The results are presented in Table 1.

TABLE 1. METRIC RESULTS FOR THREE VIDEOS

| Video \ Metric: | FP* | FN | GT | IDS | MOTP | MOTA |
|---|---|---|---|---|---|---|
| Motinasmultifront(1269 frames) | 0 | 1004 | 4219 | 20 | 0.63 | 0.757 |
| Motinasmultiwild (1007 frames) | 0 | 1015 | 3321 | 41 | 0.58 | 0.682 |
| Motinasfast (487 frames) | 0 | 314 | 1179 | 35 | 0.55 | 0.704 |

*The use of the LTM removes all false positives

The most interesting fact is the absence of false positives in all videos. This is due the use of the LTM for filtering false positives in the re-detection stage, as described in section 3.

We briefly describe each metric used in the table. MOTA (*Multiple Object Tracking Accuracy*) is the most widely used for tracking performance evaluation. It combines false negatives (FN), false positives (FP) and mismatches or identity switches (IDS) in one formula (4.1.). false negatives, false positives and identity switches (IDS) are summed together across all frames and divided with the number of ground truth (GT) detections across a video. Since the number of false positives can surpass all other values, MOTA can range between -∞ and 1.

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDS_t)}{\sum_t GT_t} \quad (4.1.)$$

MOTP (*Multiple Object Tracking Precision*) metrics measures a total error in the estimated position of the matched targets averaged by the total matches in the entire video (4.2.). In formula (4.2.) $d_{t,i}$ is overlap truth (Intersection over Union) distance measure between a bounding box and assigned ground in each frame, while $c_t$ determines the number of matchings. MOTP ranges between 0.5 and 1. It is only precision metric for localisation that ignores false positives, false negatives and ground truth relations.

$$MOTP = \frac{\sum_{t,i} d_{t,i}}{\sum_t c_t} \quad (4.2.)$$

IDS (*Identity switches*) counts how many switches in objects labelling happened during the tracking when comparing to the ground truth.

Some tracking results can be seen in Fig. 4.

Fig. 4. Tracking results.

## V. CONCLUSION

This work has presented a system for robust online multi-face tracking in video sequences recorded by a stationary camera.

The system combined several components and algorithms such as the NPD face detector, multiple DSST trackers, a PSR-based failure detector and a Hungarian trajectory assignment algorithm. The FIFO LTM/STM for storing the position correlation filters of each DSST tracker and FIFO TM for storing trajectory data for the DSST trackers were used to resolve the problem of re-detection ambiguity. The preliminary experimental results show that the system is quite robust and suitable for real world applications.

## ACKNOWLEDGMENT

This work has been supported by the Croatian Science Foundation under project 6733 De-identification for Privacy Protection in Surveillance Systems (DePPSS).

## REFERENCES

[1] W. Luo, J. Xing, A. Milan, X. Zhang, W. Liu, X. Zhao and T.K. Kim,"Multiple Object Tracking: A Literature Review", https://arxiv.org/abs/1409.7618, 2017.

[2] W. L. Lu, J. A. Ting, J. J. Little and K. Murphy," Learning to track and identify sportplayers from broadcasrt sports videos", IEEE PAMI, vol. 35, pp. 1704-1716, 2013.

[3] D. Koller, J. Weber and J. Malik",Robust multiple car tracking with occlusion reasoning", IEEE ECCV, pp. 189-196, 1994.

[4] Z. Khan, T. Balch and F. Dellaert,"An MCMC-based particle filter for tracking multiple interacting targets", IEEE ECCV, pp. 279-290, 2004.

[5] H. Pirsiavash, D. Ramanan and C. C. Fowlkes, "Globally optimal greedy algorithms for tracking a variable number of objects," IEEE CVPR, pp. 1201-1208. 2011.

[6] H. Jiang, S. Fels and J.J. Little,"A linear programming approach for multiple object tracking", IEEE CVPR, pp.1-8, 2007.

[7] L. Zhang, Y. Li and R. Nevatia, "Global data association for multi-object tracking using network flows", IEEE CVPR, pp.1-8, 2008.

[8] A. R. Zamir, A. Dehghan and M. Shah, "GMCP-Tracker: Global multi-object tracking using generalized minimum clique graphs.", IEEE ECCV, vol. 2, pp. 343–356., 2012.

[9] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little and D. G. Lowe, " A boosted particle filter: Multitarget detection and tracking", IEEE ECCV, vol.1, pp.28-39, 2004.

[10] T. E. Fortmann, Y. Bar-Shalom and M. Scheffe, " Multi-target tracking using joint probabilistic data association", IEEE Conference on Decision and Control including the Symposium on Adaptive Processes", vol. 19, pp. 807-812, 1980.

[11] S. Oh, S. Russell and S. Sastry,"Markov Chain Monte Carlo Data Association for Multi-target Tracking", IEEE Transactions on Automatic Control, vol. 54, pp. 481-492, 2009.
[12] J. Munkres, "Algorithms for the assignment and transportation problems", Journal of the society for industrial and applied mathematics, vol. 5, pp.32-38, 1957.
[13] A. Milan, K. Schindler and Stefan Roth, "Multi-Target Tracking by Discrete -Continuous Energy Minimization. ", IEEE PAMI vol 38., pp. 2054-2068, 2016.
[14] L. Zhang and L. van der Maaten," Structure preserving object tracking", IEEE CVPR, pp. 1838-1845, 2013.
[15] W. Hu, X. Li, W. Luo, X. Zhang, S. Maybank and Z. Zhang,"Single and multiple object tracking using log-euclidean Riemannian subspace and block-division appearance model", IEEE PAMI, vol.34, pp.2420-2440, 2012.
[16] S. Liao, A.K. Jain and S. Z. Li, "A Fast and Accurate Unconstrained Face Detector", IEEE TPAMI, vol. 38, Issue:2, 2016, pp. 211-223.
[17] M. Danelljan, G. Häger, F. S. Khan and M.Felsberg, "Accurate Scale Estimation for Robust Visual Tracking", Proceedings of the British Machine Vision Conference (BMVC), pp.1-11,2014.
[18] D. S.Bolme, J. R. Beveridge, B. A. Draper and Y. M. Lui, " Visual object tracking using adaptive correlation filters", IEEE CVPR, pp.2544-2550, 2010.
[19] P. F. Felzenszwalb, R. B. Girschick, D. A. McAllester and D. Ramanan, " Object Detection with discriminatively trained partbased models", IEEE TPAMI, vol.32, no.9, pp.1627-1645, 2010.
[20] M. Kristan, R. Pflugfelder, A. Leonardis, J. Matas, L. Čehovin, G. Nebehay, T. Vojir, and G. Fernandez et al., "The visual object tracking vot2014 challenge results," In IEEE ECCV Workshops, 2014.
[21] S. Zhang, Y. Gong, J. B. Huang, J. Lim, J. Wang, N. Ahuja and M. H. Yang," Tracking persons of interest via adaptive discriminative features", IEEE ECCV, pp.415-433., 2016.
[22] M. Soldić, D. Marčetić, M. Maračić, D. Mihalić and S. Ribarić, "Real-Time Face Tracking under Long-Term Full Occlusions", IEEE ISPA, pp.1-6, 2017.
[23] B. Keni and S. Rainer, "Evaluating multiple object tracking perfomance:the clear mot metrics", EURASIP Journal on Image Video Process, 10 pages, 2008.
[24] ftp://motinas.elec.qmul.ac.uk/pub/multi_face/