Multi-Target Tracking Using Hierarchical Convolutional Features and Motion Cues

Heba Mahgoub, Khaled Mostafa, Khaled T. Wassif, Ibrahim Farag
Faculty of Computers and Information
Cairo University
Cairo, Egypt

***Abstract*—In this paper, the problem of multi-target tracking with single camera in complex scenes is addressed. A new approach is proposed for multi-target tracking problem that learns from hierarchy of convolution features. First fast Region-based Convolutional Neutral Networks is trained to detect pedestrian in each frame. Then cooperate it with correlation filter tracker which learns target's appearance from pretrained convolutional neural networks. Correlation filter learns from middle and last convolutional layers to enhances targets localization. However correlation filters fail in case of targets full occlusion. This lead to separated tracklets (mini-trajectories) problem. So a post processing step is added to link separated tracklets with minimum-cost network flow. A cost function is used, that depends on motion cues in associating short tracklets. Experimental results on MOT2015 benchmark show that the proposed approach produce comparable result against state-of-the-art approaches. It shows an increase 4.5 % in multiple object tracking accuracy. Also mostly tracked targets is 12.9% vs 7.5% against state-of-the-art minimum-cost network flow tracker.**

Keywords*—*Multi-target tracking; correlation filters; convolution neural networks

I. Introduction

Multi-target tracking task is to estimate number of targets and their trajectories across multiple frames. It is a crucial problem in the field of computer vision. Also it is highly demanded in many computer application such as surveillance, human behavior analysis and augmented reality. Mainly it consists of two components: detection and data association between detections across frames. Data association step is challenging due to many reasons such as missed or faulty detections, short and long term occlusions and interactions between targets in crowded scenes. Most recent approaches in multi-target tracking have followed tracking-by-detection approach, where object detectors output are linked to build targets trajectories.

Recently, convolutional neural networks (CNN) have gained a lot of attention. CNN demonstrated the state of the results in various computer vision tasks such as object recognition, semantic segmentation and object detection. Due to it's ability to capture a generic feature representation from visual data. However, CNN rarely used in multi-target tracking. As CNN require collecting large number of training positive and negative samples, which is not always available. Also dealing with ambiguity in the decision boundary between positive and negative samples. As sampling is done near target which lead to high correlation between positive and negative samples.

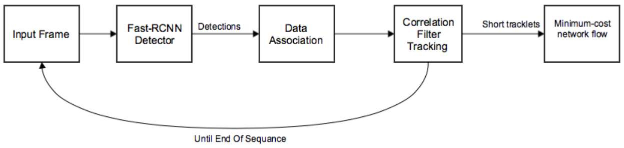

Fig. 1. Overview of out multi-target tracking system.

On the other hand CNN learned features have outperformed hand crafted features in many vision problems. As stated in previous work in single object tracking task [1] features from last convolutional layer encode target semantic information and more robust to handle appearance change in the target. However they have low spatial details which is necessary in target localization task. On the other hand features in earlier layers have high spatial details. So it's more helpful in localization but less invariant to target appearance change.

In this paper Fast Region-based Convolutional Neutral Networks (Fast RCNN) detector is integrated with correlation filters tracker. The proposed multi-target tracking algorithm is based on correlation filters that learn from hierarchical convolutional features of a pretrained CNN. Also cosine similarity is used between convolutional features along with Euclidean distance as association measure between previous frame tracklets and current detection. Then a post processing step is added with minimum-cost network flow tracker to recover target after long occlusions. Overview of our proposed approach is shown in Fig. 1.

The following three contributions are made. First Fast RCNN is integrated with correlation filters tracker. Using Fast RCNN to detect pedestrians in each frame. Also Fast RCNN is cooperated with correlation filters to handle targets disappearance. Second new data association metric is proposed between detections and trajectories that describe the path of target instances over time. Data association is based on measuring cosine similarity between middle convolutional layers. Third a fix for occlusion problem in correlation filters is proposed. Using min-cost network flow to avoid short tracklets and high identity switch rate.

This paper is organized as follow: In Section II, previous work is discussed. In Section III, The proposed approach details are discussed including basic idea of Fast R-CNN, mathematical concept for correlation filters tracker, data association metrics strategy and min-cost network flow. In Section

IV, evaluation of proposed approach on MOT2015 benchmark is presented. Comparison against state-of-the-art approaches is shown. In Section V, advantages, limitations of proposed approach and future work are discussed. In Section VI, conclusion for the work done in the paper is presented.

II. Related Work

A. Object Detection

Recently deep convolutional neural networks have made a huge progress in object detection. RCNN [2] is a detector that classify proposal regions with a deep convolutional neural networks. RCNN first compute region proposals using separated algorithm such as selective search [3]. Then it feeds the candidate regions to convolution neural networks to classify selected regions. However RCNN is slow as it doesn't share computation while performing forward pass. It process each proposal separately. Girshick proposed Fast RCNN [4] an end-to-end architecture with shared convolutional layer. Fast RCNN improved train and test speed while also gained higher detection quality.

B. Multi-target Tracking

Due to the importance of multi-target tracking in computer vision, a large number of sophisticated approaches have been developed to handle this challenging task. Specially in case of crowded scene where occlusion and false positive are common. Most work in multi target tracking follow tracking by detection approach [5, 6, 7], where it can be divided into two steps. First detect all targets in each frame. Then link theses detections to form trajectories. Processing detections can be done online, where only past and current frames are considered in building tracklets. Different approaches have been proposed in handling data association in online tracking. Early approaches handled data association by using recursive Bayesian filters such as: Kalman filter[8], Particle filter [9] which depends on first-order Markov assumption. Another direction in association is to match between objects at consecutive frames using similarity measure, where only local features are considered such as object appearance, distance between detections and size. However local association that considers consecutive frame have limitation in handling false positive and missed detections.

On the other hand other approaches in multi-target tracking adopt batch learning approach [10, 11, 12, 13, 14], where future detections are also considered and data association construct targets trajectory globally. The Association between detections is then formulated as minimization of cost function. Data association problem can be formulated to achieve global optimum by using linear programming relaxation [10, 15] or minimum-cost flow [11, 16, 17].

C. Deep Learning Multi-Target Tracking

Recent multi-target tracking algorithms based on CNN [18] or Recurrent Neural Networks have been proposed. They show higher performance when compared to handcrafted features. A Siamese CNN [18] was used to estimate likelihood if two pedestrian belong to same entity using images and optical flow as model input. Then they used gradient boosting to combine features from Siamese CNN features with contextual features. An end-to-end learning with Long Short-Term Memory (LSTM) was proposed in [19] for online tracking. Although this work was the first fully end-to-end learning method based on deep learning. Its performance did not achieve the accuracy of the state-of-the-art methods.

D. Tracking with Correlation Filters

Another recent approach in tracking is based on correlation filters. It starts with a cropped image of the tracker from a given position. After initialization, from every new frame an image patch is cropped from the estimated position. Features is extracted from the cropped image and a cosine function is applied for smoothing the discontinuities at boundary. Afterwards a correlation is computed between input and trained filter in frequency domain. Then apply inverse Fourier transform on correlation to get confidence map which give high values at the estimated target position and low values to the background. These filters can be considered as simple linear classifiers.

A new approach was proposed in [20], where all translated samples collected from target will be used in training the classifier. This enhanced training performance, without sacrificing much speed. Enhancement was done by exploiting circular structure of the kernel matrix. Extended Kernel Correlation Filter was proposed in [21] using both depth and color features. Also depth distribution was used to identify scale changes and reflect these changes in the Fourier domain. Depth was used to detect occlusion based on checking if there are sudden changes in target depth histogram. This approach achieved real time performance as it work on 35 fps. A fast Scalable Kernel Correlation Filter was introduced in [22]. This approach used Gaussian window function to deal with fixed size limitation in the kernelized correlation filter. So it allowed target scale changes and provide better separation around the target.
An extensive survey on correlation filters is available at [23], where experiments have been conducted on correlation filters to evaluate the effectiveness and efficiency of different algorithms.

III. Proposed Algorithm

Taking inspiration from previous approaches in multi-target tracking. The proposed approach is subdivided into two modules: multi-target detection and two step tracking. First step in tracking is based on correlation filters tracker for each target in scene. Each correlation filters learn from hierarchically of convolution features. Second step minimum-cost network flow is applied to link short tracklets from previous step. Our goal is to obtain entire trajectory for each target in the scene. Also each target will be associated with unique ID. Overview of our approach is shown in Fig. 2. Algorithm 1 summarize the proposed approach.

A. Fine-Tune Multi-Target Detector

An end-to-end multi-object detector is adopted which called Fast RCNN. Fast RCNN was trained on PASCAL VOC dataset. Since 2DMOT2015 benchmark is based on pedestrian. A fine-tune step is applied to consider pedestrians only. So softmax layer is changed to only consider two classes: pedestrians and non pedestrians.

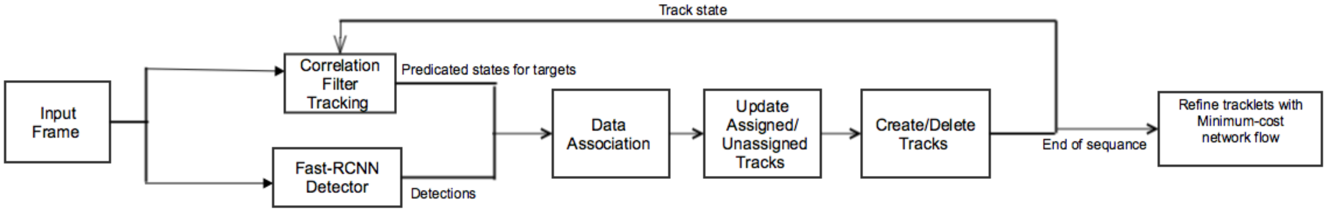
Track state
Correlation Filter Tracking
Predicated states for targets
Input Frame
Data Association
Update Assigned/ Unassigned Tracks
Create/Delete Tracks
End of sequance
Refine tracklets with Minimum-cost network flow
Fast-RCNN Detector
Detections

Fig. 2. Detailed description for our system using correlation filer and minimum-cost network flow.

A selective search algorithm is used to generate region proposals for the network. Object detector will measure each proposal region and give each detection a score. Only detections with a score higher than a threshold will be considered as valid. High predefined threshold = 0.9 will cause most pedestrians to remain undetected. However decreasing detection threshold will cause an increase in false positive, which is more severe than missed detections. Final step non-maximal suppression is applied based on bounding box overlap between detections in order to suppress redundant boxes.

B. Correlation Filters Tracking

As mentioned above. Many possible targets states remain undetected. Correlation filters tracking is applied to improve detection. Correlation filters (CF) have attracted a lot of attention in recent years for speed and accuracy. Due to analyzing frames in Fourier domain which lead to faster processing. It has the ability to update appearance model at every frame. CF is a discriminative classifier that learns to separate target form it's background. The main idea behind CF tracker is that a learned filter is used to predict target position by searching maximum value in correlation response map. We follow the same mathematical model in model learning [1]. CF learns from features that were extracted from VGG-Net-19 [24] which was trained on imageNet [25]. The proposed algorithm in [1] used output from conv 3-4, conv 4-4 and conv 5-4. Due to pooling operation in VGG-net-19 which cause gradual decrease in spatial resolution. This lead to imprecise target localization. For example convolution feature size of pool4 is 14x14 and in pool5 is 7x7. In order to solve this problem Convolution feature is resized to fixed larger size using bilinear interpolation.

A correlation filter W^l is learned from each convolution map to generate response map, where l indicate number of convolutional maps. Feature vector of l-layer of size MxNxD is denoted as x , where M,N and D indicate width,length and height. Output from applying Gaussian function to the circular shifts of x along the M and N dimensions is denoted as y. Each W^l is updated in frame t from previous frame t-1 using the numerator A^l and the denominator B^l through the following equations:

$$A_t^l = (1-\eta)A_{t-1}^l + \eta Y \odot \overline{X}_t^d; \qquad (1)$$

$$B_t^l = (1-\eta)B_{t-1}^l + \eta \sum_{i=1}^{D} X_t^i \odot \overline{X}_t^i; \qquad (2)$$

$$W_t^l = \frac{A_t^d}{B_t^d + \lambda} \qquad (3)$$

The capital letters refer to Fourier transformed signals. $\odot$ indicate element wise multiplication and η is a learning rate and λ is a regularization parameter.

The a correlation response map is calculated given new image patch that contain target with the following equation:

$$F^l = IFFT(\sum_{d=1}^{D} W^d \odot \overline{Z}^d); \qquad (4)$$

where IFFT symbol for inverse Fast Fourier transform .Then target new position can be deduced by searching maximum value of correlation response map.

C. Detection Guidance

We propose cooperation between Fast RCNN detector and CF tracker to handle CF disadvantages such as: scale variation and model drifting problem. As CF doesn't handle scale variation well. We add CF predicted bounding box of targets to region proposals of Fast RCNN. Then use predicted scales to updated targets appearance model. This way the detector will validate CF predictions. Also we use Fast RCNN detector to discover if target disappeared from point of view. We consider detector score, if it's less than predefined threshold. We know that target state is inactive and stop update it's appearance model. So this step will prevent model drift that may occur to the tracker. As shown in Fig. 3 and 4.

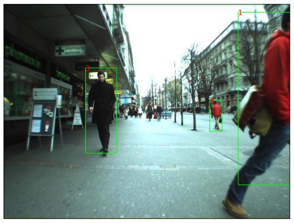

Fig. 3. Frame i-1.

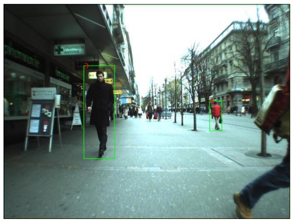

Fig. 4. Frame i.

Example of detection guidance in our model. At frame i-1 we have three targets each assigned an ID and target with ID '1' appear to be leaving the view. At frame i the target '1' disappears. So bounding box disappears.

D. Data Association

Data association goal is to associate between current frame detections and tracks that describe targets paths. This lead to updating each target state and identify the new detections. Each target can belong to a single state. Target state can be one of the following: assigned, unassigned, lost and new.

Hungarian assignment algorithm [26] is used to achieve this task. According to the assignment algorithm results we can determine the following: assigned detections with tracks, unassigned tracks and new detection. In each frame we apply Hungarian algorithm and use cosine similarity as cost measure between current frame detections and predicted target position from CF. Highly overlapped bounding boxes are considered in the association. The cosine similarly is calculated between middle convolution features from output of conv 3-4 layer. Since the middle layer has higher spatial detail which is useful in differentiating between targets. The overlap function between two bounding boxes A,B is calculated as follows:

$$Overlap = \frac{A \cap B}{A \cup B} \tag{5}$$

E. Minimum-Cost Network Flow

This is done as post processing step after the proposed multi-target correlation filters tracking. The integration between Fast RCNN with correlation filters handle low missed detection rate and increase tracker precision. Also our proposed approach can handle some cases of occlusion such as targets occluded by non pedestrian objects or leaving scene. On the other hand we still need to handle recovering target after long occlusion. Also we need to handle the case where target is occluded by other pedestrian. All these issues may cause wrong association and false target model update. We follow batch learning approach to handle these issues, where we can use future detections. Multi-target tracking is formulated with minimum-cost network flow. Matching between detections is solved jointly for all tracklets.

We used modified version of minimum-cost network flow in [12] to refine the resulted tracklets. As it depends on motion cues in the association between detections. Given initial set of tracklets that consists of a set of ordered detections. Motion cues are used to refine those tracklets and link separated short tracklets.

For every detection d in tracklet. Two sets of detections is defined. The first set contains all tracklet detections before d. The second set contains all tracklet detections after d. These two sets are used to determine linear regression coefficients that can predict forward and backward target position. Then cost function that considers residual between the predicted and actual tracklet positions is computed. Finally a minimum-cost network flow solution is computed to produce the final tracklets.

Algorithm 1: Pseudocode of our Proposed Approach

```
INPUT: Video
OUTPUT: tracklets + bounding box with ID
for each frame in video do
    Detect pedestrian using Fast RCNN
    Apply non-maximal suppression
    for each correlation filters tracker do
        Update target location from proposed regions
    end for
    Data Assignment based on updates states
    for each detection in unassigned detection do
        Initialize new correlation filters tracklet
    end for
    for each tracklet in assigned tracklets do
        Update target model from convolutions maps
    end for
    for each tracklet in unassigned tracklets do
        Run Fast RCNN to the predicted location

        if detector confidence = Threshold then
            Update target model from convolutions maps
        else
            Delete tracklet
        end if
    end for
end for
for each tracklet do
    Refine tracklets using motion cue LP optimization
end for
```

IV. Experiments

In the evaluation step 2DMOT2015 [29] benchmark is used to evaluate our multi-target tracking algorithm. A common reference in multi-target tracking task. 2DMOT2015 benchmark is composed of training and testing sets. Training set consists of a 11 sequence with a 40,000 bounding box, while testing set consists of 11 sequence with 60,000 boxes. 2DMOT2015 benchmark contains sequences with high target motion variation, camera motion, a different views and person density. Also for fair comparison in tracking task, 2DMOT2015 provide public detections, given by Aggregate Channel Features (ACF) pedestrian detector [30]. In order to be able to compare the proposed work with others, public detections are used as region proposals for the Fast RCNN detector.

The widely accepted CLEARMOT evaluation metrics [31] are employed by 2DMOT2015. To summarize multi-object tracking (MOT) performance the following measures were reported: MOT accuracy (MOTA) measures jointly three errors: false positives (FP), false negatives (FN) and identity switches (IDSwitch). MOT precision (MOTP) measures the misalignment between the detected target locations and ground truth. Also mostly tracked and mostly lost targets percentages (MT and ML) are reported. Furthermore, the IDSwitch ratio between targets is reported.

A. Analysis on MOT Validation Data

In order to train Fast RCNN detector to detect pedestrians. Training data in 2DMOT2015 benchmark are split into two parts to fine-tune our detector: train and validation. Training data include the following sequences: TUD-Stadtmitte, ETH-Bahnhof, ADL-Rundle-6, PETS09-S2L1 and KITTI-13. While testing data include TUD-Campus, ETH-Sunnyday, ETH-Pedcross2, ADL-Rundle-8, KITTI-17 and Venice-2

TABLE I. COMPARISON WITH STATE-OF-THE-ART APPROACHES. THE BEST SCORE ARE BOLDFACED. ARROW UP INDICATE HIGHER IS BETTER. WHILE ARROW DOWN INDICATE LOWER IS BETTER

Algorithm	MOTA ↑	MOTP ↑	MT ↑	ML ↓	FP ↓	FN ↓	IDSwitch ↓
oICF [27]	27.1	70.0	6.4 %	48.7 %	7,594	36,757	**454**
TBX [28]	27.5	70.6	10.4 %	45.8 %	7,968	35,810	759
RNN_LSTM [19]	19.0	71.0	5.5 %	45.6 %	11,578	36,706	1,490
Siamese CNN [18]	29.0	71.2	8.5 %	48.4 %	**5,160**	37,798	639
ELP [12] (Baseline)	25.0	71.2	7.5 %	43.8 %	7,345	37,344	1,396
Ours multi CF	25.5	71.9	9.3 %	**34.4** %	12,344	31,378	2,064
Ours multi CF+minimum-cost n/w	**29.5**	**73.1**	**12.9** %	36.3 %	11,866	**30,474**	976

B. Evaluation on MOT Testing Data

In order to be fair in comparing with other approaches. During testing phase, public detections are used which were provided by benchmark as region proposals for Fast RCNN detector. So Fast RCNN will only filter 2DMOT2015 public detections and eliminate false positive. **Baseline comparison** The proposed approach is compared with minimum-cost network flow in [12]. To show the progress achieved by the proposed multi-target tracking algorithm in improving precision and restoring undetected targets states.

The proposed approach is compared against state-of-the art deep learning based approach such as Siamese CNN in [18] and recurrent neural network in [19], as shown in Table.I. Also the proposed approach is compared against approaches based on handcrafted features in [27, 28].

The results show that cooperating Fast RCNN with multi-correlation filters tracker produce high precision and low missed detection rate. Also the benefit of using motion cue with minimum-cost network lowered identity switch which improve the mostly tracked targets rate.

V. DISCUSSION

Training correlation filters with hierarchy of convolution features improves tracker robustness and accuracy. Also co-operating correlation filters with Fast RCNN helps tracker from drifting and scale estimation problem. These lead to low missed detection rate and high tracker precision. The last step in the proposed approach is refining the tracklets while considering motion similarity. However refining tracklets with linear velocity assumption may fail in case of random motion patterns, which lead to false association and increase identity switch between targets. Non linear motion patterns will be considered in future work.

VI. CONCLUSION

In this paper, multi-target tracking algorithm is proposed that exploit features from pretrained convolutional neural network. First Fast RCNN is trained to detect all pedestrians in the scene. Then a correlation filters tracker is proposed to learn target appearance. It learns from hierarchy of convolution features. As middle convolution layers are useful for target localization while last convolutional layer are more robust in handling target appearance changes. Also cosine similarity is used between convolution features in data assignment between tracklets and detections.

Finally to handle correlation filters failure in case of occlusion a minimum cost network is proposed to link short tracklets. Experimental results demonstrate that the proposed algorithm provides competitive performance on the 2DMOT2015 benchmark.

REFERENCES

[1] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang, “Hierarchical convolutional features for visual tracking,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3074–3082.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[3] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, “Selective search for object recognition,” *International journal of computer vision*, vol. 104, no. 2, pp. 154–171, 2013.

[4] R. Girshick, “Fast r-cnn,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[5] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, “Robust tracking-by-detection using a detector confidence particle filter,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 1515–1522.

[6] W. Choi, “Near-online multi-target tracking with aggregated local flow descriptor,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3029–3037.

[7] C.-H. Kuo, C. Huang, and R. Nevatia, “Multi-target tracking by on-line learned discriminative appearance models,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 685–692.

[8] R. E. Kalman *et al.*, “A new approach to linear filtering and prediction problems,” *Journal of basic Engineering*, vol. 82, no. 1, pp. 35–45, 1960.

[9] K. Okuma, A. Taleghani, N. De Freitas, J. J. Little, and D. G. Lowe, “A boosted particle filter: Multitarget detection and tracking,” in *European conference on computer vision*. Springer, 2004, pp. 28–39.

[10] H. Jiang, S. Fels, and J. J. Little, “A linear programming approach for multiple object tracking,” in *Computer Vision and Pattern Recognition, 2007. CVPR’07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[11] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[12] N. McLaughlin, J. M. Del Rincon, and P. Miller, “Enhancing linear programming with motion modeling for multi-target tracking,” in *Applications of Computer Vision (WACV), 2015 IEEE Winter Conference on*. IEEE, 2015, pp. 71–77.

[13] H. Pirsiavash, D. Ramanan, and C. C. Fowlkes, “Globally-optimal greedy algorithms for tracking a variable number of objects,” in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 1201–1208.

[14] A. Milan, K. Schindler, and S. Roth, “Multi-target tracking by discrete-continuous energy minimization,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 10, pp. 2054–2068, 2016.

[15] Z. Wu, A. Thangali, S. Sclaroff, and M. Betke, “Coupling detection and data association for multiple object tracking,” in *Computer Vision and*

Pattern Recognition (CVPR), 2012 IEEE Conference on. IEEE, 2012, pp. 1948–1955.

[16] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, "Multiple object tracking using k-shortest paths optimization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 33, no. 9, pp. 1806–1819, 2011.

[17] P. Lenz, A. Geiger, and R. Urtasun, "Followme: Efficient online min-cost flow tracking with bounded memory and computation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4364–4372.

[18] L. Leal-Taixé, C. Canton-Ferrer, and K. Schindler, "Learning by tracking: Siamese cnn for robust target association," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 33–40.

[19] A. Milan, S. H. Rezatofighi, A. R. Dick, I. D. Reid, and K. Schindler, "Online multi-target tracking using recurrent neural networks." in *AAAI*, 2017, pp. 4225–4232.

[20] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista, "Exploiting the circulant structure of tracking-by-detection with kernels," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 702–715.

[21] M. Camplani, S. L. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, and T. Burghardt, "Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling." in *BMVC*, 2015, pp. 145–1.

[22] J. Lang, R. Lagani *et al.*, "Scalable kernel correlation filter with sparse feature integration," in *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*. IEEE, 2015, pp. 587–594.

[23] Z. Chen, Z. Hong, and D. Tao, "An experimental survey on correlation filter-based tracking," *arXiv preprint arXiv:1509.05520*, 2015.

[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[25] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[26] J. Munkres, "Algorithms for the assignment and transportation problems," *Journal of the society for industrial and applied mathematics*, vol. 5, no. 1, pp. 32–38, 1957.

[27] H. Kieritz, S. Becker, W. Hübner, and M. Arens, "Online multi-person tracking using integral channel features," in *Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*. IEEE, 2016, pp. 122–130.

[28] R. Henschel, L. Leal-Taixé, B. Rosenhahn, and K. Schindler, "Tracking with multi-level features," *arXiv preprint arXiv:1607.07304*, 2016.

[29] L. Leal-Taixé, A. Milan, I. Reid, S. Roth, and K. Schindler, "Motchallenge 2015: Towards a benchmark for multi-target tracking," *arXiv preprint arXiv:1504.01942*, 2015.

[30] P. Dollár, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[31] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, no. 1, p. 246309, 2008.