Multi Target Tracking by Linking Tracklets with a Convolutional Neural Network

Yosra Dorai[1,2], Frederic Chausse[1], Sami Gazzah[2] and Najoua Essoukri Ben Amara[2]
[1]*Institut Pascal, Blaise Pascal University, Clermont-Ferrand, France*
[2]*LATIS, Laboratory of Advanced Technology and Intelligent Systems, ENISo, Sousse University, Sousse, Tunsia*
{*yosra.dorai, sami.gazzah*}*@gmail.com, frederic.chausse@univ-bpclermont.fr, najoua.benamara@eniso.rnu.tn*

Keywords: Multi-object Tracking, Tracklet, Faster R-CNN, Traffic Surveillance, Occlusion.

Abstract: The computer vision community has developed many multi-object tracking methods in various fields. The focus is put on traffic scenes and video-surveillance applications where tracking object features are challenging. Indeed, in these particular applications, objects can be partially or totally occluded and can appear differently. Usual detection methods generally fail to leverage those limitations. To deal with this, a framework for multi-object tracking based on the linking of tracklets (mini-trajectories) is proposed. Despite the number of errors (false positives or missing detections) made by the Faster R-CNN detector, short-term Faster R-CNN detection similarities are tracked. The goal is to get tracklets in a given number of frames. We suggest to associate tracklets and apply an update function to correct the trajectories. The experiments show that on the one hand, our approach outperforms the detector to find the undetected objects. And on the other hand, the developed method eliminates the false positives and shows the effectiveness of tracking.

1 INTRODUCTION

The tracking is the estimation of the possible trajectories of an object as it moves in a scene. Its goal is that every object keeps the same label despite occlusion, similarity, and detector defects. In fact, there are three steps in tracking: locating positions, estimating the motion of object and following its movement. Various applications use tracking to analyse scenarios in the domain of autonomous driving, visual surveillance and robot navigation.

Several methods of object tracking exist in the literature as: tracking by detecting (Badie, 2015), feature-based (Hadi et al., 2014), and 3D-model-based (Battini and Landi, 2015). A lot of object-tracking methods have been proposed not only to cover a variety of viewpoints like object poses, backgrounds, lighting conditions..., but also to keep track of object identities over time in spite of frequent occlusion by clutter or other objects and similar appearances of different objects. Recently, on multi-object tracking, most of the works have used the tracking-by-detection strategy and the data association of detected objects. The basis of the algorithms is to research the similarities between detected objects. Some cues are combined to compute similarity like appearance, location and movement...

In this paper, we put forward a method of tracking using data association based on the tracking approach. It links short track fragments (tracklets) and detection responses into trajectories by global optimization. In the first step, we utilize the Faster R-CNN detector (Simonyan and Zisserman, 2014) (the first time used on tracking according to our knowledge). In the second step, tracking combines the detection responses to build initial tracklets. In our approach, we opt for a global stage to associate detections. We are looking at the similarity of objects from appearance, position and speed. In addition, we give each object a specific signature. We has also contribute by adding an update stage to correct the trajectories.

The paper is organized as follows. We start with a presention of the related works in section 2. After that, our approach is described in section 3. Next, the experiments are detailed in section 4. Finally, we conclude and present the future work in section 5.

2 RELATED WORKS

As we have already mentioned, we have focused on the method of tracking by detection (composed by detection and tracking steps).We give a brief review of

Dorai Y., Chausse F., Gazzah S. and Essoukri Ben Amara N.
Multi Target Tracking by Linking Tracklets with a Convolutional Neural Network.
DOI: 10.5220/0006155204920498
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 492-498
ISBN: 978-989-758-227-1

the main detectors and the major tracking methods.

Detectors. The recent detectors have witnessed a significant progress. In what follows, we choose to review some of the important detectors. Sevral authors have chosen to use the Histogram of Oriented Gradient (HOG) to detect objects (Badie and Bremond, 2014). For example, (Mao and Yin, 2015) utilize the HOG to detect pedestrians. Recently, object detection from still images has been mainly based on deep learning (Szegedy et al., 2015) (Girshick, 2015) (Ouyang et al., 2015) (Gidaris and Komodakis, 2015). The literature has shown that the neural networks and essentially the Faster R-CNN outperform other detectors (Mao and Yin, 2015)(Szegedy et al., 2015). Although the HOG and deep learning are often used for detection, they have a significant number of false positives and missed ones. That is why some researchers (Mao and Yin, 2015)(Maamatou et al., 2015) have opted for a specialized method to increase the performances of the detector.

Tracking. The literature of multi-object tracking is vast. However, we can divide the tracking methods into two categories: on-line and off-line. We are interested in the on-line or recursive method (Erdem et al., 2004) applied in real time. It uses on-line information and is based only on past observation to build the trajectory. Therefore, it is more difficult to deal with the missing detection. As a result, it will have a lot of trajectories for the same object.

Serveral works use two steps to associate detections: the local and global associations. Furthermore, the local one considers a few frames to solve the association problem. Instance, in (Bar-Shalom et al., 1980), despite the ambiguities in association, they used the association probabilities that would compute across all targets. In a global association, the number of frames is important and can be the entire of video (Dehghan et al., 2015) (Zamir et al., 2012). Indeed, in most of the results, the trajectory is not complete and can be fragmented to many trajectories for the same object. Thus, why some researchers have utilize the method of an associated short trajectory (known as tracklet) (Yang and Nevatia, 2012) (Bae and Yoon, 2014). To link tracklets several works were based on appearance (Kuo et al., 2010; Nillius et al., 2006) and motion (Yang and Nevatia, 2012). The previous works used in matching mini trajectories the Hungarian method (Bae and Yoon, 2014) or the linear programming (Erdem et al., 2004). There is no rule in selecting the matching method; each work of the literature uses a different method.

In this paper, we address the challenging problems in long-term tracking of multiple objects in a complex scene captured by a single camera. We have chosen to work with the method of tracking by detection. We have also chosen to use the Faster RCNN detector for detection and the association of the mini trajectories for tracking.

3 MULTI-OBJECT TRACKING FRAMEWORK

3.1 Approach Overview

Our approach is inspired by (Bae and Yoon, 2014). Indeed, the authors categorize a tracklet based on their degree of confidence. They use both local and global associations. We reformulate the architecture of association to eliminate the tracklet classification. This is intended to understate the steps of association, increase the performance and manage the occlusion problems. We add an update step to correct the tracklet from the previous step: If the tracklet contains holes in a certain frame from the beginning to the end of its construction. There, we will check if a detection was not attributed to a tracklet. We are also inspired by the approach from (Mao and Yin, 2015). In fact, they used the tracklets to increase the detection performances. Accordingly, we choose to work with a tracklet to solve the problems of tracking independently from correcting the detector defects. In other words, our approach is composed by 4 steps (fig.1). The first step is the input which can be a video (I). The next step is the detection by the Faster R-CNN (II). Then, in the step of tracking (III) we construct the tracklet. We associate detection to have initial tracklets. In addition, we compare them with detection from the regressor function of the Faster R-CNN. We also associate the tracklets which have almost the same signatures (contain the characteristics of each object). We update the tracklets by adding non associated detections. Finally, output step contains the trajectory of each object with a unique ID (IV).

3.2 Detection by Faster R-CNN

The detector Faster R-CNN is composed of two modules. The first one is Region Proposal Network (RPN) which is deep fully convolutional network that serves to provide regions. The second one is the Fast R-CNN. It has been proven that the Faster R-CNN is faster than the previous versions (RCNN, SPPNET and Fast R-CNN). We have exploited the model of caffe VGG 16 of (Simonyan and Zisserman, 2014) with 13 convolutional layers. The RPN is modeled by a fully convolutional network. Its input is an image.

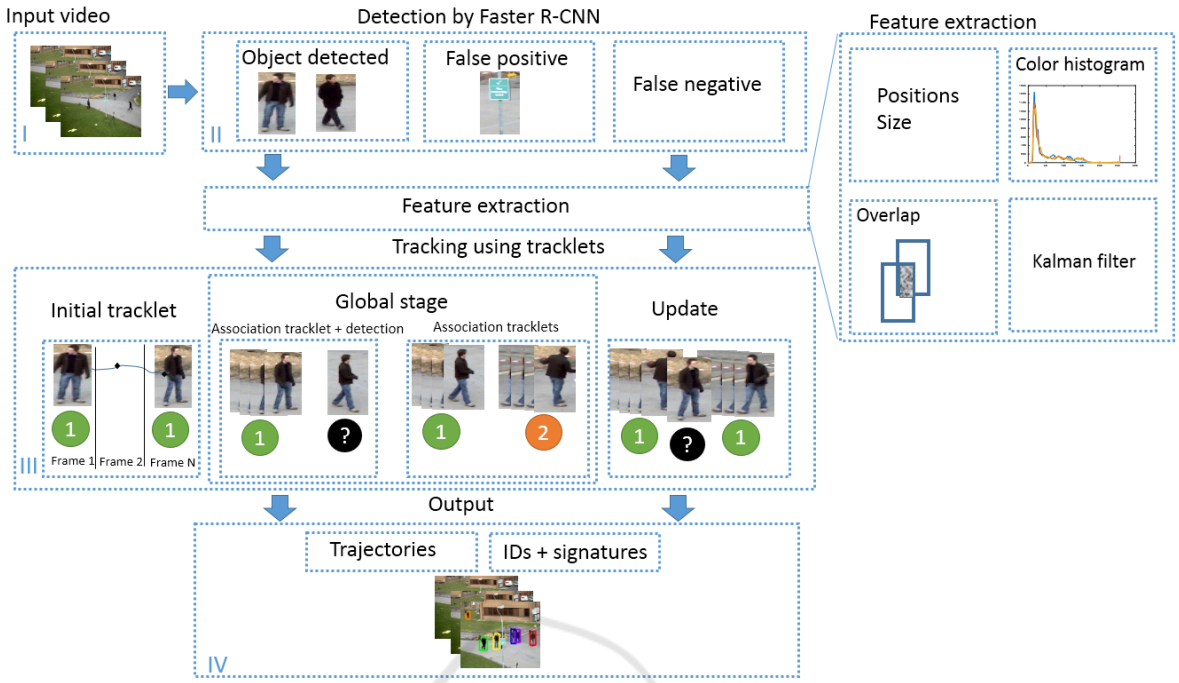

Figure 1: Block diagram of our tracking method.

It is an $n * n$ spatial windows of the output of the convolutional feature map by the last convolutional layers. The movement of the sliding windows provides a vector. It is characterized by a lower dimension and it is introduced in the fully connected, box regression and box classifier layers. The second module of the Faster R-CNN is the Fast R-CNN. It takes as an input the output of the RPN. The network is composed by convolutional and max-pooling layers to produce a convolutional feature map. However, to use the Fast R-CNN in the Faster R-CNN, the convolutional layers will be shared by the Fast R-CNN and the RPN. We will use in this paper the generic Faster R-CNN using the RPN and the Fast R-CNN and we are essentially interested in the box regressor function.

3.3 Tracking using Tracklets

The advantages of the detection by the Faster R-CNN is that it naturally identifies new objects of interest entering the scene. This detector presents false positives and missing detection but we consider these problems minor and negligible because with a tracklet we can predict and correct these defects.

3.3.1 Tracklet

A tracklet is a chain of nodes O^i representing one single detected object which appears in N frames with the same ID i from the start time t_s^i to the final time t_f^i. Each node O_t^i is one detection at the time t ($t \in [t_s^i, t_f^i]$). It has a unique signature defined by features (e.g. localization, speed, size, appearance...). One object can have multiple chains in one scene because of missed detection or occlusion problems. Indeed, when such problem occurs, the initial tracklet is interrupted. Then, when the same object appears again, later a new tracklet is initiated instead of continuing the ancient one. Hence, tracklets have to be associated to solve occlusion problems and to predict missed objets. The association is made by the Hungarian method (Ahuja et al., 1993).

3.3.2 Initial Tracklets

Initial tracklets are built just after the detection step. First,we choose the number N of frames on which the tracklets are defined. We associate the detection from t_s to t_f. The association is done according to the overlap and similarity of appearance between successive detections. After initial tracklet constructions, we can predict the following positions using the Kalman filtering (Chong et al., 2014).

3.3.3 Global Stage

In this stage, we associate tracklets having similar signatures as well as detections provided by the system. Each detected object passes through the extraction feature block to define their characteristics. Most

Algorithm 1: Our approach of tracking.

Require: Video or frame sets
The number of frame to construct tracklet N
Tracklet for the object i : T^i
Node of tracklet of the object i : O^i.
The detection by Faster R-CNN : $detection$.
Ensure: trajectories

$k \leftarrow 0$
repeat
`/* intial tracklets */`
for k to N **do**
Construct intial tracklets
end for
`/* Global step */`
Provide the next detection from intial tracklet
Compare them with the output of the regressor function of Faster R-CNN
if $detection \neq 0$ and and corresponds to the estimate **then**
Association detections
end if
Compare tracklet
if signature T^i and signature T^j are resembled **then**
Association tracklet
end if
`/* update step */`
while $detection$ not associated **do**
Compare $detection$ with Nearest tracklets
if there is a resemblance **then**
Update tracklet
end if
end while
until all objects are successfully tracked

Most methods adopt affinity measures to compare two detections across time, such as special affinity (e.g. bounding box overlap, Euclidean distance or simple appearance similarities). The advantages of our method is to solve crossing and occlusion problems. To associate detections we use the overlap level and we check similarity at the same time. In this way, if two objects are detected, we will check the overlap between objects in frames F_{n-1} and F_n and their similarity.

Nevertheless, we can find problems (e.g. ID switch, tracklets of the same object or merging two or more tracklets). This is maybe caused by the cross of two objects, occluded by a background element or behavior of an object. We present thereafter every problem and how to remedy it.

First, we start with the behavior of objects. Indeed, each object is characterized by bounding box dimensions, a trajectory (direction and speed: determined by the Kalman filter), and a characteristic vector that contains appearance information determined by a color histogram (HSV). The object can undergo a natural phenomena in a scene such that the object can leave scene. To solve this problem, we validate the next hypothesis: If we have more than four missing nodes in a tracklet, we do not attribute the ID anymore. If we follow a single object, the tracking will be made by the overlap between the detected object in frame F_n and the object detected in frame F_{n-1}, beside recording the features of the object at each node of the tracklet. Here we use the box of a regressors' function of the Faster R-CNN because it is adjusted and precise on the object in order to build the initial tracklets. From initial tracklets, we provide the detections in the next frame and we compare the prediction with box of regressor function.

Second, we treat the case of several objects having an overlap. We can get this when the objects intersect or approach : We compare the appearance to infer classes.

Third, we treat the case of occlusion between an object and its environment as follows: We have in this case frames with undetected objects and subsequently missing nodes in the tracklets or two tracklets of the same object with different IDs. To remedy the problem of missing nodes, prediction is performed using the features of the two nodes located just before and just after the missing nodes. Then, we can have more than two missing nodes. Therefore, we have more than one tracklet for one object with different IDs. To solve this, we compare in our global stage, the signatures of the tracklets. In fact, a signature contains the features of appearance, position, size and speed of each tracklet node. If the features of two traklets are similar, we associate the two tracklets by a Hungarian method. If not we attribute for each tracklet an ID.

3.3.4 Update

In the global stage, if the algorithm misses detection in frame F_t and is not able to do a correct matching, then we can provide the missing one, during the update step. This is explained by a defect in the construction of the tracklet. In other word, it means that some nodes of the tracklet are missing. Indeed, the tracklet can not be built in the case when the number of missing nodes is superior to the number of frames N (the number of frames to define the tracklet). Also, the number of missing nodes in a tracklet must be less than four. In fact, by this strategy, we can eliminate false positives. On the other hand, if the number of missed nodes is more than four frames, we will consider that the object is out of scene and that the ID is

Table 1: CLEAR MOT metrics tracking results on the monocamera sequence PETS09 S2L2, PETS09 S2L1 and ETHMS (Sunny and Bahnhof).

Dataset \ Metrics		Presion	Recall	MOTA	MOTP	FP
PETS S2L1	(Breitenstein et al., 2011)	-[1]	-	0.79	0.56	-
	(Bae and Yoon, 2014)	-	-	0.83	0.69	0.19
	Our approach	**0.93**	**0.82**	**0.86**	**0.69**	**0.05**
PETS S2L2	(Poiesi et al., 2013)	-	-	0.59	-	-
	(Bae and Yoon, 2014)	-	-	0.7	0.53	0.14
	Our approach	**0.99**	**0.6**	**0.68**	**0.54**	**0.002**
ETHMS (Sunny and Bahnhof)	(Bae and Yoon, 2014)	-	-	0.72	0.64	0.04
	Our approach	**0.9**	**0.7**	**0.74**	**0.66**	**0.01**

not used any longer. In order to update the missing one, we check non-associated detection and we compare the object appearances with the signatures of the closest tracklets. The main goal of this update step is to correct the tracklet, add non associated detection and track objects as much as possible. Finally, we obtain in the output the trajectory of each object with its unique ID and signature.

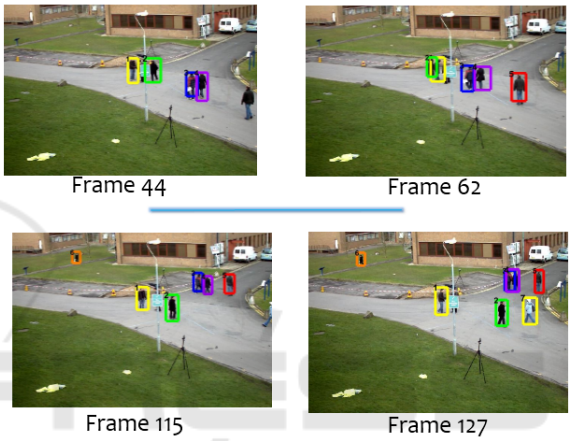

Figure 2: Tracking results: IDs (1 yellow, 2 green) correctly kept despite crossing.

4 EXPERIMENTS

In this section, we present the used metrics and the obtained results.

4.1 Metrics

For all the PETS2009 S2L1, PETS 2009 S2L2 and ETHMS (Sunny and Bahnhof) sequences, we use the CLEAR MOT metrics (MOTP, MOTA, presion, recall and false positives (FP)) in order to compare with related works. The MOTP metric is multiple-object tracking precision to evaluate the tracking results with bounding boxes of ground truth. The MOTA metric is the multiple-object tracking accuracy to mesure the ID switch, the false positives and the false negatives.

4.2 Results

Table [1] gives the results of our tracking method and a selection of the works of the state of the art. Our tracker is evaluated on the PETS2009 S2L1 dataset. The efficiency of our approach is observed in fig.2, since the two pedestrians (1 yellow and 2 green) have kept their ID from frame 44 to frame 127, despite the presence of a crossing problem. Our approach is able to improve the results by increasing the metric. The tracker is also evaluated on the PETS2009 S2L2 dataset and ETHMS (Sunny and Bahnhof). For this dataset, we also use the metric (MOTA, MOTP, presion, recall and false positives (FP)).

By analyzing Table 1, we notice that the number of false positives is less than other works. This is thanks to the use of a high performance detector and to our technique neglecting the false positives. In addition, our MOTA results are encouraging because this metric depends on the ID switch. Actually, we have managed to keep the ID for each object as long as possible.

5 CONCLUSION

In this paper, we have proposed a new approach of tracking using a tracklet based on the function box regressor of Faser R-CNN. The framework is evaluated on the public datasets PETS2009 (S2L1 and S2L2) and ETHMS (Sunny and Bahnhof). The tracking based on a tracklet can solve the problems of occlusion by providing the missing detection. The

[1]No results found.

global approach based on the Faster R-CNN detection presents a reduced number of false positive trajectories. We have successfully found a way to keep the same ID for each object as long as possible by our "update step" to correct tracklets and associate non detections.

In the future work, we will use a network camera to track objects.

ACKNOWLEDGEMENT

This work is sponsored by a co-guardianship between the University of Sousse (Tunisia) and Blaise Pascal University (France).

REFERENCES

Ahuja, R. K., Magnanti, T. L., and Orlin, J. B. (1993). Network flows: theory, algorithms, and applications.

Badie, J. (2015). *Optimizing Process for Tracking People in video-camera network*. PhD thesis, Universite Nice Sophia Antipolis.

Badie, J. and Bremond, F. (2014). Global tracker: an online evaluation framework to improve tracking quality. In *Advanced Video and Signal Based Surveillance (AVSS), 2014 11th IEEE International Conference on*, pages 25–30. IEEE.

Bae, S.-H. and Yoon, K.-J. (2014). Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1218–1225. IEEE.

Bar-Shalom, Y., Fortmann, T., and Scheffe, M. (1980). Joint probabilistic data association for multiple targets in clutter. In *Proc. Conf. on Information Sciences and Systems*, pages 404–409.

Battini, C. and Landi, G. (2015). 3d tracking based augmented reality for cultural heritage data management. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 40(5):375.

Breitenstein, M. D., Reichlin, F., Leibe, B., Koller-Meier, E., and Van Gool, L. (2011). Online multiperson tracking-by-detection from a single, uncalibrated camera. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1820–1833.

Chong, C.-Y., Mori, S., Govaers, F., and Koch, W. (2014). Comparison of tracklet fusion and distributed kalman filter for track fusion. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–8. IEEE.

Dehghan, A., Modiri Assari, S., and Shah, M. (2015). Gmmcp tracker: Globally optimal generalized maximum multi clique problem for multiple object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4091–4099.

Erdem, C. E., Sankur, B., and Tekalp, A. M. (2004). Performance measures for video object segmentation and tracking. *IEEE Transactions on Image Processing*, 13(7):937–951.

Gidaris, S. and Komodakis, N. (2015). Object detection via a multi-region and semantic segmentation-aware cnn model. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1134–1142.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1440–1448.

Hadi, R. A., Sulong, G., and George, L. E. (2014). Vehicle detection and tracking techniques: a concise review. *arXiv preprint arXiv:1410.5894*.

Kuo, C.-H., Huang, C., and Nevatia, R. (2010). Multi-target tracking by on-line learned discriminative appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 685–692. IEEE.

Maamatou, H., Chateau, T., Gazzah, S., Goyat, Y., and Amara, N. E. B. (2015). Transfert d'apprentissage par un filtre séquentiel de monte carlo: application à la spécialisation d'un détecteur de piétons. In *Journées francophones des jeunes chercheurs en vision par ordinateur*.

Mao, Y. and Yin, Z. (2015). Training a scene-specific pedestrian detector using tracklets. In *2015 IEEE Winter Conference on Applications of Computer Vision*, pages 170–176. IEEE.

Nillius, P., Sullivan, J., and Carlsson, S. (2006). Multi-target tracking-linking identities using bayesian network inference. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 2187–2194. IEEE.

Ouyang, W., Wang, X., Zeng, X., Qiu, S., Luo, P., Tian, Y., Li, H., Yang, S., Wang, Z., Loy, C.-C., et al. (2015). Deepid-net: Deformable deep convolutional neural networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2403–2412.

Poiesi, F., Mazzon, R., and Cavallaro, A. (2013). Multi-target tracking on confidence maps: An application to people tracking. *Computer Vision and Image Understanding*, 117(10):1257–1272.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Stauffer, C. (2003). Estimating tracking sources and sinks. In *Computer Vision and Pattern Recognition Workshop, 2003. CVPRW'03. Conference on*, volume 4, pages 35–35. IEEE.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Tran, Q.-V. (2016). Non-contact breath motion detection using the lucas-kanade algorithm.

Xing, J., Ai, H., and Lao, S. (2009). Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1200–1207. IEEE.

Yang, B. and Nevatia, R. (2012). Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1918–1925. IEEE.

Zamir, A. R., Dehghan, A., and Shah, M. (2012). Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In *Computer Vision–ECCV 2012*, pages 343–356. Springer.