What makes for good multiple object trackers?

Yuqi Zhang, Yongzhen Huang, Liang Wang
Center for Research on Intelligent Perception and Computing
National Laboratory of Pattern Recognition,Institute of Automation,Chinese Academy of Sciences
Beijing, China
{yuqi.zhang, yzhuang, wangliang}@nlpr.ia.ac.cn

***Abstract*—This paper explores the importance of detection and appearance features for multiple object tracking. Extensive detectors including hand-crafted methods and deep learning methods have been tested. We found in this paper that simply improving detection performance can lead to much better multiple object tracking results. The data association methods used in this paper are Kalman Filter and Hungarian algorithm as proposed in [1]. CNN features and color histogram features are extracted as appearance features to measure similarities between objects. Our experiments show that appearance features can help with data association. We then combine detection and data association together as an overall system. The proposed system can track multiple objects at a speed of 17 fps with high accuracy.**

***Index Terms*—Deep learning; Multiple Object Tracking; Appearance Features**

I. Introduction

Multiple Object Tracking (MOT) is important in many applications such as visual surveillance and intelligent robots. After we have recognized people in videos, the next step is to keep track of them whenever they are occluded or move out of screens. This avoids the heavy computation of recognizing people all the time.

Tracking can be divided into two types: single object tracking and multiple object tracking. One obvious difference is the number of objects. The other is whether we need to give an ID to each object. Single object tracking can be seen as a way to speed up detection. It does not care about a person's ID. On the other hand, multiple object tracking should give an ID to each person. The aim is to deal with occlusion, merge and split problems. In this way, multiple object tracking is also regarded as a data association problem. In this paper, we only focus on multiple object tracking.

Current multiple object tracking benchmarks such as MOT 2015 [2] focus on the tracking-by-detection principal. These benchmarks often provide detection results for each frame. Different algorithms are performed on the same detection results from Aggregate Channel Features (ACF) [3] pedestrian detector. The algorithms should assign each detection with an ID or a false alarm. There are mainly two types of methods: batch-based methods and online methods. Batch-based methods use future frames to correct current misleading association while online methods can only use history frames. Classic batch-based data association methods including Multiple Hypothesis Tracking (MHT) [4] and Joint Probabilistic Data Association(JPDA) [5] have been revisted these years. The authors claimed that these classic methods could also perform well on the recent benchmarks. However, these methods are batch-based and run slowly. Thus they are not suitable for real-time applications.

For real-time applications such as robot navigation and autonomous driving where we cannot use future information, online methods might be useful. In contrast to many other complicated methods, Bewley et.al in [1] ranked top at the MOT 2015 benchmark with very classic methods Kalman Filter [6] and Hungarian algorithm [7]. Ignoring appearance features, it only used the position and size of the bounding box for both motion estimation and data association. It also ignored short-term and long-term occlusion to make the system simple and fast. The sufficiently simple data association method performed well on the MOT 2015 benchmark mainly because of the better detection results by Faster RCNN detector [8].

Deep learning methods have been widely used in many computer vision fields such as classification, localization and object detection [9, 10, 8]. Recently single object tracking has witnessed great improvement with the use of deep learning [11, 12]. However, for multiple object tracking deep learning methods are not widely used. This is the motivation of our paper. Note that we are different from Bewley et.al in [1]. First, they only used Faster RCNN detectors trained on VOC dataset which has a class of person. VOC dataset is quite different from MOT 2015 dataset in both the objects and the scenes. Models trained on VOC may not generalize well. On the contrast, we train pedestrian specific detectors on MOT 2015 dataset. Second, appearance features were ignored by Bewley et.al while we take CNN features and color histogram features of the objects into consideration. Last, we combine the detector with the data association algorithm to achieve real-time multiple object tracking.

This paper has the following three main contributions:

- Several detectors are tested to evaluate the importance of detection to multiple object tracking.
- CNN features and color histogram features are extracted to provide more information to data association.
- A real-time deep learning based system is proposed for multiple object tracking.

The paper is organized as follows. Section II compares different detection methods including ACF and CNN detectors. Section III analyzes different options for appearance features. Section IV describes our real-time deep learning based multiple object tracking system while Section V concludes the paper.

978-1-5090-4165-7/16/$31.00 ©2016 IEEE

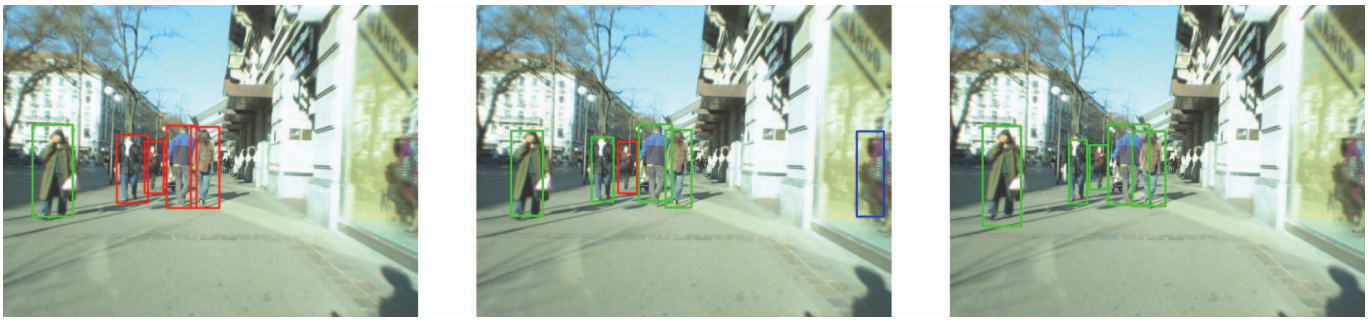

Fig. 1: Comparison between different detectors. ACF(left), VGG16 trained on VOC(middle) and VGG16 trained on MOT 2015 (right). The green rectangles, red rectangles and blue rectangles stand for detected objects, missed objects and false alarms respectively.

II. Detection

This section will discuss several detectors including ACF and Faster RCNN. We will use the same data association algorithm as [1] and try different detectors.

A. Detector

The ACF detector uses hand-crafted features: normalized gradient magnitude, histogram of oriented gradients (6 channels), and LUV color channels. The features are then fed into decision trees to distinguish pedestrians from background. This method is fast enough to run at 30 fps on a single PC. The MOT 2015 benchmark uses the ACF detector trained on INRIA dataset to detect people at each frame.

Faster RCNN [8] is a deep learning detection framework that proves to be very successful. It first generates proposals by a Region Proposal Network and then classifies the objects in the region proposals. Though it is originally designed for VOC object detection, it is also successful on other datasets. With the help of modern GPUs, Faster RCNN could reach a speed of about 18 fps.

B. Finetuning

Different from [1] which used the Faster RCNN model trained on VOC dataset, we fintune our model on MOT 2015 dataset. We emphasize that the people and the scenes of the two datasets are different. Note that MOT 2015 takes occlusion into consideration. That is to say, when a person is totaly occluded, he still has a groundtruth bounding box. Also the evaluation method still counts the unseen person. That is the reason why the detectors have a low recall. When we train our model, we should not use the occluded groundtruth even though they are given by the dataset. In detail, we first erase out groundtruth bounding boxes that are totally occluded by others. We do not ignore partly occluded groundtruth bounding boxes as the detector should have the ability to detect them. We use two convolutional neural networks: ZF [10] and VGG16 [13]. We train a total of 70,000 batches with a batch size of 128. The learning rate is initially set to 0.001 and is decreased by 10 after 50,000 batches.

TABLE I: Evaluation of detection and tracking performance on the same validation set as [14].

Detector	recall	precision	IDs	MOTA
ACF [1]	33.6	65.7	224	15.1
ZF [1]	41.3	72.4	347	24.0
VGG16 [1]	49.5	77.5	274	34.0
proposed ZF(0.5)	33.4	87.7	200	27.8
proposed VGG16(0.5)	38.5	95.1	174	35.7
proposed ZF(0.3)	35.8	84.9	211	28.5
proposed VGG16(0.3)	40.0	93.9	180	36.7

C. Performance Evaluation

The MOT 2015 benchmark collects widely used video sequences for multiple object tracking. These sequences are divided into a training set and a test set each with 11 sequences. Since the annotations of the test set are not released, a validation set of 6 sequences from the 11 training sequences is seperated to give further analysis [14]. We use the standard CLEAR MOT [15] metrics to evaluate our models. Identity switches (IDs) mean the number of times an ID switches to a different previously tracked object. A model that can keep the ID of a pedestrian unchanged after occlusion should have low IDs. The multi-object tracking accuracy (MOTA) combines errors such as false positives (FP), false negatives (FN) and identity switches (IDs) into a single number. Recall and precision are also provided.

Figure 1 illustrates detection results of ETH-Sunnyday sequence frame 1. Three detectors including ACF, VGG16 trained on VOC and VGG16 trained on MOT 2015 are listed. As can be seen, the original ACF detector performs badly. We should not expect the data algorithm to bring the lost pedestrians back. Our model trained on MOT 2015 performs a little better, with fewer false alarms and more detected objects. This is because our model has seen similar scenes and pedestrians in the training sequence ETH-Bahnhof.

Table I shows the tracking results of different detectors. CNN detectors are better than ACF detectors. As for the same CNN architecture, the same confidence threshold of 0.5 is used. MOTA indicates that the models trained on MOT

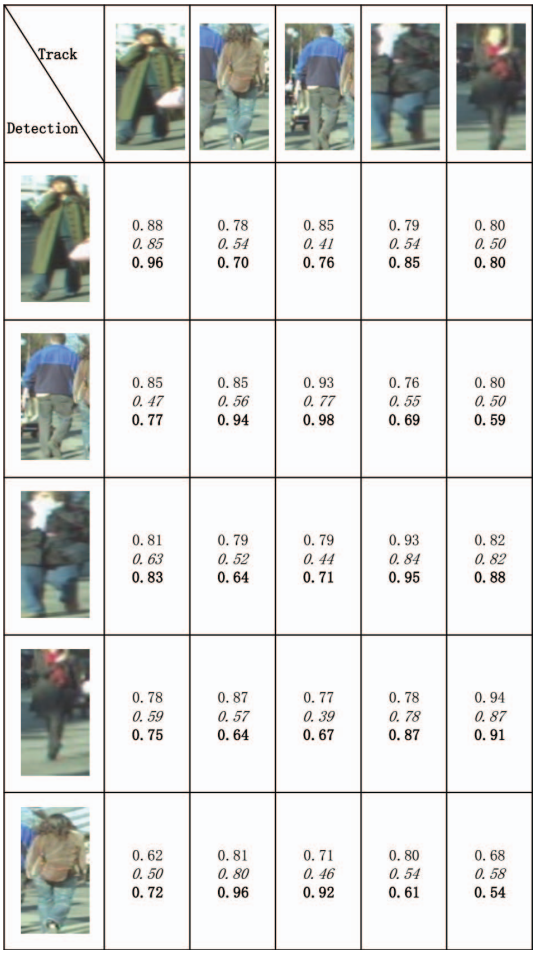

Fig. 2: Cosine distances between detections and tracks with three different features. VGG16 detector(normal font), VGG16 Imagenet(italics) and color histogram(bold).

2015 are better. Note that we use the same data association algorithm, which means the better results come from the better detectors. We also found on the validation set that the confidence threshold could influence the performance. Models with a threshold of 0.3 outperform models with a threshold of 0.5.

III. Appearance features

In this section, we will extract appearance features by CNN and color histogram. The same detection results from VGG16(0.5) are used to make sure the only difference is the appearance representation. Cosine distances between detections and tracks are calculated to measure the similarities between them. The motivation is that objects with similar appearances should be matched together.

The CNN features are extracted from the last fully connected layer with a dimension of 4096. We have two sources of CNN features: Imagenet pretrained and MOT 2015 detection pretrained in Section II. Which model should we use? Qualitatively, Imagenet pretrained model has the ability to distinguish 1000 classes, which means it can focus on details including shape, color, texture and so on. Since there is no ZF network available, we use caffe reference Alexnet [16] and VGG16 pretrained on Imagenet. Each bounding box region is warped to a fixed size (227×227 for Alexnet and 224×224 for VGG16) before the feature extraction. On the other hand, the model trained on MOT 2015 only classifies people and background. It does not care much about the differences between two people. This model might not capture enough information to describe the appearance of people. For the CNN detector features, we directly output fc7 features of Faster RCNN without warping.

TABLE II: Performance evaluation when only considering appearance features.

Appearance Feature	recall	precision	IDs	MOTA
VGG16 detector	28.6	70.2	1194	11.3
Alexnet Imagenet	38.1	91.0	248	33.2
VGG16 Imagenet	38.1	91.6	269	33.4
Color histogram	38.2	91.5	233	33.7

Classic color histogram is also used to extract appearance features. For a RGB image, we compute the color histogram of 256 bins. In total we have a feature vector of 768 dimension for an object. This is relatively small compared with the 4096 dimension of CNN features.

Figure 2 lists the cosine similarities between detections and tracks in ETH-Sunnyday frame 1 and frame 2. It can be seen that detector features give close similarities for almost all the objects. This means detector features are not suitable for feature representation. VGG16 Imagenet model gives much better similarities. Detection 3 has close similarities to Track 4 and Track 5 because of blur. For color histogram features, Detection 5 is regarded similar to Track 2 and Track 3. However, they are not similar in appearance.

In [1], intersection-over-union (IOU) is computed as the distance between detections and tracks. IOU only focuses on the position and ignores the appearance. On the contrary, we ignore the position and focus on the appearance. Cosine distances between detections and tracks are computed to replace the IOU distances. This means objects with similar appearance should be matched together.

Table II illustrates the performance when we only consider appearance features. Imagenet pretrained model is better than MOT detection pretrained model, which proves our qualitative analysis. We found that people in some sequences have very similar appearance. Appearance features can no longer help for these sequences. Interestingly, the simple color histogram features outperform VGG16 Imagenet features. One possible explanation is that VGG16 Imagenet features are extracted from warped images which might lose a lot of information.

Since we have position and appearance information, we combine them together to provide more information to the data association algorithm. Most of the time, we only use position information. When occlusion happens, position information might fail. In [1], tracks are removed when they are unseen for

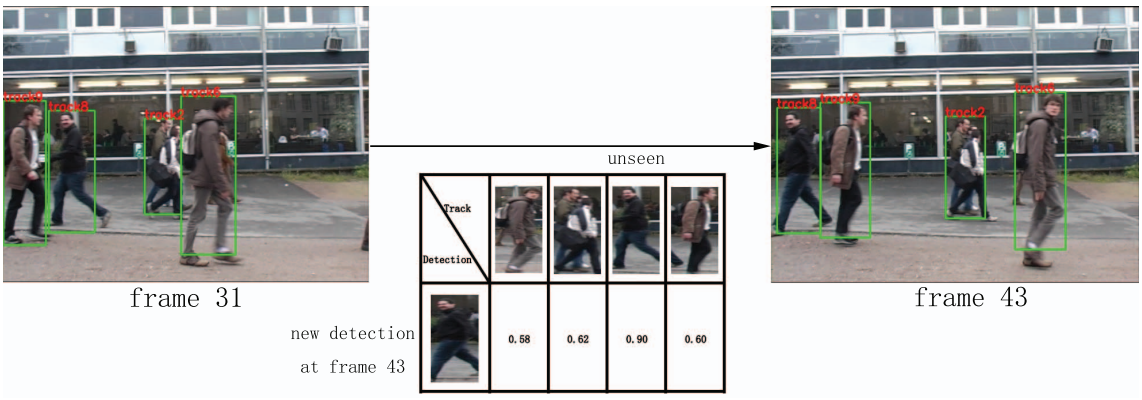

Fig. 3: The framework with appearance features to help data association.

TABLE III: performance evaluation when combining position and appearance together.

Appearance Feature	recall	precision	IDs	MOTA
position(no appearance)	38.6	94.7	107	36.0
VGG16 detector	38.8	94.3	106	36.0
Alexnet Imagenet	38.8	94.5	114	36.0
VGG16 Imagenet	38.9	94.3	106	36.1
Color histogram	38.6	94.4	118	35.9

more than one frame. We enable longer lives of tracks by tolerating unseen tracks for at most 20 frames. Figure 3 illustrates the case. When one person (track 8) is totally occluded during frame 31 to frame 43, he can still have his ID remained by Kalman Filter for at most 20 frames. When he is detected again at frame 43, [1] is likely to give a new ID to him because of inaccurate Kalman Filter prediction. Appearance information can help at this time. We save the history appearance features of each person. When an unmatched detection occurs, we compute its appearance similarities with the unseen tracks. If the most similar unseen track has a similarity of over 0.5 with this unmatched detection, then the new detected box should be assigned to the unseen track.

Table III shows performance evaluation when combining position and appearance together. It can be seen that the model with the combination of position and VGG16 Imagenet appearance features works a little better. However the difference is minor. We can also find that the appearance features from different sources have similar performances. Even detection features which cannot describe appearance well give similar results. On average, the algorithm described above for the occlusion case happens only 9 times for a sequence. Note that Kalman Filter can also succeed for some of the cases. This is the reason why all methods have similar performances.

When we use the same detection results from the proposed VGG16 model, the position only data association algorithm outperforms the appearance only one (MOTA 35.7 vs 33.4). Also when we allow longer lives of tracks, the data association algorithm with the combination of position and appearance features does not improve much (MOTA 36.1 vs 36.0). We can conclude that detection quality is more important than appearance features.

IV. Real-time System

Previous multiple object tracking algorithms often focus on high performance rather than speed. This is not suitable for real-time applicants. In the above sections, we found that detection contributes a lot to the overall performance and CNN appearance features can help a little when occlusion happens. In this section, we will propose a real-time multiple object tracking system. The system takes raw video frames as input and outputs bounding boxes with IDs for each frame. The ID of a person could be kept when occlusion happens.

As the data association algorithm is fast enough to run at over 400 fps, the main bottleneck of the system is the detection part. We choose ZF network in Section II, which runs at 18 fps on a NVIDIA Titan X. We do not use CNN appearance features because extracting these features takes extra time. Combined with Kalman Filter and Hungarian Algorithm, the whole system runs at 17 fps on 640×480 videos. Note that we also keep a considerable tracking performance with a MOTA of 28.5 on the validation set.

V. Conclusion

In this paper, we found that detection contributes a lot to multiple object tracking performance. It turns out that training on MOT 2015 dataset gives better detection performance and thus better tracking performance. We also include CNN features and color histogram features in data association. Experiments show that good appearance features can help data association. Last, a real-time multiple object tracking system is proposed. Since we still divide multiple object tracking into detection and data association, the next step would be an end-to-end multiple object tracking system.

References

[1] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft, "Simple online and realtime tracking," in *arXiv preprint arXiv:1602.00763*, 2016.

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

[3] P. Dollar, R. Appel, S. Belongie, and P. Perona, "Fast feature pyramids for object detection," *Pattern Analysis and Machine Intelligence*, vol.36, 2014.

[4] C. Kim, F. Li, A. Ciptadi, and J. M. Rehg, "Multiple hypothesis tracking revisited," in *International Conference on Computer Vision*, 2015, pp. 4696–4704.

[5] S. H. Rezatofighi, A. Milan, Z. Zhang, A. Dick, Q. Shi, and I. Reid, "Joint probabilistic data association revisited," in *International Conference on Computer Vision*, 2015, pp. 3047–3055.

[6] R. Kalman, "A new approach to linear filtering and prediction problems," *Journal of Basic Engineering*, vol. 82, 35-45, 1960.

[7] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol.2, 83-97, 1955.

[8] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 91–99.

[9] 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.

[10] M. Zeiler D and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV 2014*, 2014, pp. 818–833.

[11] L. Wang, W. Ouyang, X. Wang, and H. Lu, "Visual tracking with fully convolutional networks," in *International Conference on Computer Vision*, 2015, pp. 3119–3127.

[12] C. Ma, J. Huang, and M. Yang, "Hierarchical convolutional features for visual tracking," in *International Conference on Computer Vision*, 2015, pp. 3074–3082.

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

[14] Y. Xiang, A. Alahi, and S. Savarese, "Learning to track: Online multi-object tracking by decision making," in *International Conference on Computer Vision*, 2015, pp. 4705–4713.

[15] K. Bernardin and R. Stiefelhagen, "Evaluating multiple object tracking performance: The clear mot metrics," *Image and Video Processing*, May, 2008.

[16] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*, 2014, pp. 675–678.