Deep Person Re-identification for Probabilistic Data Association in Multiple Pedestrian Tracking

Brian H. Wang\textsuperscript{1}, Yan Wang\textsuperscript{2}, Kilian Q. Weinberger\textsuperscript{2}, and Mark Campbell\textsuperscript{1}

Abstract—We present a data association method for vision-based multiple pedestrian tracking, using deep convolutional features to distinguish between different people based on their appearances. These re-identification (re-ID) features are learned such that they are invariant to transformations such as rotation, translation, and changes in the background, allowing consistent identification of a pedestrian moving through a scene. We incorporate re-ID features into a general data association likelihood model for multiple person tracking, experimentally validate this model by using it to perform tracking in two evaluation video sequences, and examine the performance improvements gained as compared to several baseline approaches. Our results demonstrate that using deep person re-ID for data association greatly improves tracking robustness to challenges such as occlusions and path crossings.

I. INTRODUCTION

Visually tracking the motion of people through a scene over time is a critical capability for many applications involving camera-equipped robots or sensor networks. Examples range from an autonomous car tracking nearby pedestrians, to a team of aerial robots searching for moving people in a search-and-rescue mission. This problem can be broken down into two general stages: people must be detected in video frames, and these detections must then be linked together and used to estimate tracks over time. The first of these two tasks has been extensively studied with the usage of deep convolutional neural networks for object detection [7], [12], [6], [20]. State of the art algorithms such as Mask-RCNN [8] are capable of detecting people at the pixel level together and used to estimate tracks over time. The first of these two tasks has been extensively studied with the usage of deep convolutional neural networks for object detection [7], [12], [6], [20]. State of the art algorithms such as Mask-RCNN [8] are capable of detecting people at the pixel level with near-human level accuracy, and have been found to generalize well to different scenes.

However, for multiple pedestrian tracking applications, it is not enough to just detect the presence of people - it is equally important to distinguish between individuals and correctly associate detections with currently tracked people. A robot which confuses individuals with one another could assume an incorrect number of people in its environment in an autonomous navigation or search-and-rescue scenario, or an intelligent sensor network could lose track of a person of interest in a security task.

This problem becomes highly challenging in the presence of occlusions, crowds, and pedestrians crossing paths with one another. See Figure 1 for an example of some of the situations that complicate multiple-person tracking. A method for modeling appearance and distinguishing between different people visually could provide key information to robustly handle these challenges.

Person re-identification (re-ID) provides a promising solution to this problem [29], [5], [15], [25]. Given two images that each contain a person, a re-ID system can generate a likelihood that the two images are of the same person. Similar to people detection in a scene, re-ID is extremely intuitive for humans - consider how quickly most people can spot a friend in a crowd, or pick out a cameo appearance by a favorite celebrity in a movie. However, re-ID is challenging for a computer, and is therefore an active area of research in machine learning and computer vision. State of the art approaches to person re-ID involve modeling the appearances of individuals in a low-dimensional feature space, that is learned with a deep convolutional neural network trained on a large dataset of images of many different people. The network is trained explicitly so that images of the same person are mapped to close-by locations in feature space, while images of different people are mapped to far apart locations.

Modern person re-ID methods are becoming very successful as measured by performance on publicly available re-ID benchmark datasets [29], [28], [1], [5], [15], [25]. However, the usefulness of deep re-ID methods for application in probabilistic tracking algorithms has not yet been extensively studied. Traditionally in pedestrian tracking, position information is used to association a new observation to a nearby tracked person. It is proposed here to additionally use appearance, via re-ID, as a cue to aid data association decisions, as seen in the example in Figure 2. This paper studies how well deep learning-based person re-identification improves data association in probabilistic multiple-person tracking.

Fig. 1. Examples of challenging situations for a multiple pedestrian tracker: a person being occluded by an obstacle, a large crowd, and two people crossing paths.
Previous images of a certain person over time steps 1, 2, ..., k

New detections of people at time step k+1

Fig. 2. Given a set of previous detections of a person, deep person re-ID could hypothetically be used to re-identify this person within the set of detections seen in a new video frame.

II. RELATED WORK

A. Probabilistic Data Association

Data association is a difficult problem in multiple-object tracking, and several proven approaches exist that achieve respectable performance. The Rao-Blackwellized particle filter (RBPF) probabilistically evaluates multiple data association decisions at each time step, and propagates multiple hypotheses of assignments forward in time [23], [24], [19]. Multiple Hypothesis Tracking (MHT) works similarly, but rather than performing probabilistic sampling, it grows a tree of possible hypotheses according to deterministic branching decisions [4], [11].

Some tracking methods perform data association over long periods of time, using probabilistic graphical models where more nodes are added to the graph as the length of time considered increases [27], [26], [10]. These methods are able to jointly reason about multiple objects in the scene over multiple time steps, and therefore generally achieve higher overall accuracy compared to online methods. However, such methods are not well-suited for applications in robotics or intelligent sensor networks, which generally require tracking to be performed recursively one frame at a time, as the system captures video data sequentially.

B. Person Re-identification

The general goal of machine learning-based person re-identification is to learn a method for mapping images of people to low-dimensional feature vectors. These feature vectors should have the property that two images which are of the same person map to nearby feature points (as measured by Euclidean distance), while images of two different people map to feature points that are far apart. This is illustrated in Figure 3. Note that ideally, re-ID feature representations should be robust to changes in the background, pose, or orientation relative to the camera, as well as to partial occlusions by obstacles.

People are able to perform re-identification by looking at visual cues such as clothing color, facial features, body shape, and distinctive accessories, to name just a few. Deep convolutional networks have been shown [7] to learn and extract high quality features from natural images, which are invariant to small local transformations such as rotation and translation. For our re-ID task, these invariants are crucial, as they help ensure that a person's identity remains unchanged as he or she moves across a scene. Indeed, many successful approaches to re-ID have used deep convolutional neural networks to map images to re-ID feature vectors [1], [5], [14], [15]. Researchers have also created a number of high-quality benchmark datasets, for the purpose of training and evaluating these methods [21], [29], [28].

C. Deep Learning for Multiple Pedestrian Tracking

Deep learning has also been applied directly to the problem of multiple pedestrian tracking. In particular, recurrent neural networks (RNNs) have been employed for tracking [17], due to their ability to effectively process time series data. Sadeghian et al. [22] use RNNs to learn motion, appearance, and interaction-based cues that indicate similarity between new observations and previously tracked pedestrians. While shown to be effective given sufficient training data, it remains to be seen whether data-driven approaches to data association and tracking are viable for robotic applications, which typically involve significant variations in environments and operating conditions, necessitating an exponentially greater amount of training data for successful performance.

III. RE-ID-BASED DATA ASSOCIATION

A. Probabilistic Data Association

In this section, we begin by defining the general data association problem within the context of multiple object tracking. Given a set of measurements $Z(k)$ and a set of tracked objects $X(k)$, where $k$ is the current time step, the goal of data association is to determine which measurement was generated by which tracked object. Individual measurements are denoted by $z_i(k) \in Z(k)$, $i = 1, \ldots, m_k$, and estimated object tracks are denoted by $x_j(k) \in X(k)$, $j = 1, \ldots, n_k$.

For the pedestrian tracking task, each time step $k$ corresponds to a distinct video frame, and the measurements $Z(k)$ are a set of bounding boxes or segmentation masks found by a computer vision person detector. Data association is then a discrete assignment problem, where each detection must be assigned to a tracked person.

This problem is shown more specifically in Figure 4. Each member of the set of new detections (Fig. 4 center) must be either assigned to a previous track (Fig. 4 left) or declared a new track (Fig. 4 right). In order to perform data association, a likelihood value is typically defined for each possible assignment from measurement.
Tracker estimate up to time step $k$
New detections observed
Tracker estimate at time step $k+1$

Fig. 4. Example of the data association problem for vision-based multiple pedestrian tracking. The left image shows estimated tracks of pedestrians up to the current time step. When a new set of detections is observed (center), each new detection must be either associated with a previously tracked person or used to initialize a new track, thereby updating the track estimates by one time step (right).

$z_i(k)$ to track $x_j(k)$, denoted as $a_{ij}(k)$. Let $\theta_{ij}(k)$ be the event that the detection $z_i(k)$ was generated by person $x_j(k)$. The likelihood values $a_{ij}(k)$ are then defined as

$$a_{ij}(k) = P(\theta_{ij}(k) | z_i(k), x_j(k))$$

(1)

for $i = 1, \ldots, m_k$ and $j = 1, \ldots, n_k$. Note that a separate likelihood model should also be defined for the event where detection $i$ is used to initialize a new track, as this falls outside the data association problem.

Equation (1) defines the probability of an assignment event only conditioned on the current detection $z_i(k)$ and track state $x_j(k)$. This is the approach used in nearest-neighbors and one-shot data association strategies [2]. Other approaches, such as the Rao-Blackwellized particle filter and MHT, operate over the full history of measurements, $Z(1), \ldots, Z(k)$, using (1) only for assignments at the last time step. Typically, even these approaches define recursive solutions, which closely examine the most recent event in (1). In this work, we focus on the individual likelihood shown in (1), but importantly, this approach can be generalized to any data association method [2], [23], [19], [4].

Since one of the events $\theta_{ij}(k)$ must explain the source of detection $z_i(k)$, and a detection cannot come from two different people, these events are mutually exclusive and exhaustive, and

$$\sum_{j=1}^{n_k} a_{ij}(k) = 1 \quad \forall i = 1, \ldots, m_k.$$  

(2)

The association likelihoods $a_{ij}$ then depend on the selection of the likelihood model $P(\theta_{ij}(k) | z_i(k), x_j(k))$ in (1). Since many of the sensors traditionally used for robotics and tracking applications provide location measurements, data association likelihoods typically depend only on the location of the sensor measurement relative to the expected object position. As an example, points returned by a lidar sensor can only be reliably associated with tracked objects according to the point positions [18]. In cases such as these, the sensor likelihood model can be accurately approximated as a Gaussian centered on the expected object position.

When using object detections in a video frame, the measurement $z_i(k)$ includes detection position as well as size (for a bounding box) or shape (for a segmenting mask). However, these detections also contain useful information about the appearance of detected people. Thus, camera detection measurements can be decomposed into two independent measurements: a position component $z_i^{\text{pos}}(k)$ and an appearance component $z_i^{\text{app}}(k)$. The measurement likelihood model then becomes

$$a_{ij}(k) = P(\theta_{ij}(k) | z_i^{\text{pos}}(k), z_i^{\text{app}}(k)).$$  

(3)

Using Bayes rule, (3) can be rewritten as

$$a_{ij}(k) = \alpha P(z_i^{\text{pos}}(k), z_i^{\text{app}}(k) | \theta_{ij}(k)) \times P(z_i^{\text{pos}}(k), z_i^{\text{app}}(k)).$$

(4)

where $\alpha$ is a normalization factor. The term $P(z_i^{\text{pos}}(k), z_i^{\text{app}}(k))$ represents any prior information that may be available about the likelihood of observing detections; it is common to simply use a uniform distribution over all possible assignments [19]. Assuming a uniform prior, and noting that the position and appearance components of the measurement are independent of one another, the assignment likelihood values are therefore

$$a_{ij}(k) = \alpha P(z_i^{\text{pos}}(k) | \theta_{ij}(k)) P(z_i^{\text{app}}(k) | \theta_{ij}(k)).$$

(5)

The likelihood of detection $z_i(k)$ being observed at a certain position $z_i^{\text{pos}}(k)$ given that it has been assigned to person $j$ can be accurately modeled as Gaussian, with the likelihood decreasing as distance to the estimated position of person $j$ increases.

Deriving the detection appearance model $P(z_i^{\text{app}}(k) | \theta_{ij}(k))$ requires augmenting the estimated state vector for each tracked person to include information on the person’s appearance, in the form of re-ID features. The augmented feature vector for person $j$ is defined as

$$\tilde{x}_j(k) = \begin{bmatrix} x_j(k) \\ f_j(k) \end{bmatrix},$$

(6)

where $f_j(k)$ is a vector in re-ID feature space. For tracking using bounding box detections, the positional component $x_j(k)$ stores the tracked person’s estimated position, velocity, and bounding box size. In our data association approach, the stored feature vector $f_j(k)$ is formed from a moving average of all feature vectors from detections assigned to person $j$. 


B. Re-ID Likelihood Model

One final question in the data association likelihood model is how to transform Euclidean distances between re-ID feature points into the likelihood values \( P(z^{opp}_i|\theta_{ij}(k)) \) used in (5). When new detections are received, they are each converted to re-ID feature vectors using a deep re-ID model. Each tracked person has an averaged reference feature vector \( f_j(k) \) associated with them.

Then, given a new detection \( z^{opp}_i(k) \), which is converted into a re-ID feature vector \( g_i(k) \), the likelihood of assigning it to person \( j \) can be calculated using the softmax function over all tracked person reference vectors, giving

\[
P_i(\theta_{ij}(k)|z^{opp}_i(k)) = \beta_i \exp \left( -\|g_i(k) - f_j(k)\| \right) .
\]

\( \beta_i \) is a normalization term used to ensure that \( \sum_{j=1}^{n_k} P(\theta_{ij}(k)|z^{opp}_i(k)) = 1 \), and therefore

\[
\beta_i = \left( \sum_{j=1}^{n_k} \exp \left( -\|g_i(k) - f_j(k)\| \right) \right)^{-1} .
\]

The softmax function is commonly used in machine learning to convert from feature vectors to class likelihoods in multiclass classification problems; since we are here interested in the minimal distance between pairs of feature vectors, we instead use a softmax to form a discrete probability distribution for \( P(z^{opp}_i|\theta_{ij}(k)) \) over the \( n_k \) possible person assignments. The full data association likelihoods shown in (5) are then finally computed by multiplying these appearance similarity-based probabilities together with the Gaussian detection position likelihoods.

C. Deep Anytime Re-ID (DaRe)

In our multiple pedestrian tracker, we use the Deep Anytime Re-ID (DaRe) architecture from Wang et al. [25] to perform person re-identification through transforming images of detected people to re-ID feature vectors. In addition to achieving state of the art performance on re-ID benchmarks, DaRe is particularly well-suited for robotic applications, because it utilizes varying amounts of computational resources depending on the person being re-identified. As an example, a person who is wearing a distinctive outfit and is clearly visible is identified using only the first stage of the DaRe convolutional neural network; someone else who is partially occluded, blurred by camera motion, or wearing clothes that blend into the background is identified using a greater number of convolutional layers.

At each stage, DaRe calculates a confidence value for the re-identification result, and stops computation when sufficiently confident. The overall effect is a significant reduction in computation time, as compared to approaches that apply computation uniformly to each person. Our implemented model is trained on the MARS dataset [28], and makes uses of a dense convolutional neural network (DenseNet) architecture [9].

IV. Experimental Validation

In order to experimentally examine the effects of incorporating appearance re-ID into a data association likelihood model, we present results from implementing the above described data association strategy within a Rao-Blackwellized particle filter. The algorithm tracks multiple moving pedestrians within two video sequences showing complex street scenes. In order to understand the impact of re-ID on data association, results are presented using four data association methods: detection position only, deep re-ID likelihood only, position along with a simple appearance model, and finally position combined with deep re-ID. Results are analyzed quantitatively using various numerical metrics of tracking performance, and qualitatively by examining specific cases where the inclusion of deep re-ID improves tracker robustness to certain difficulties.

A. Rao-Blackwellized Particle Filter

There are many different algorithms for performing data association and multiple-object tracking, based on the association likelihoods defined in (5). In order to consider uncertain data association decisions, our experiments utilize a Rao-Blackwellized particle filter (RBPF) [18], [19], which samples data association hypotheses (in the form of particles) based on the likelihood model in (5). Individual objects are tracked using efficient parametric trackers, and decisions on when to initiate new tracks from detections are also made probabilistically according to the detection association likelihoods. Detailed treatments of the RBPF can be found in Särkkä et al. [23] and Miller and Campbell [18].

Our RBPF uses separate linear motion model Kalman filters to track each individual person. Linear motion is only an approximation for the true patterns of pedestrian motion; person detectors provide frequent enough measurement updates to correct for model inaccuracy and adequately track pedestrians. The RBPF could certainly be extended with a more complex physics-based or data-driven motion model; this is outside the focus of this paper.

In order to demonstrate the applicability of deep person re-ID to multiple-pedestrian tracking, the RBPF tracker was applied to a pair of challenging video sequences taken from the Multiple Object Tracking Challenge (MOTC) benchmark [13], [16]. The PETS09-S2L1 and MOT17-04 sequences are used, shown in Figures 5 and 6 respectively. These sequences show people walking in various patterns, with occlusions by each other and by other objects, and therefore provide a useful evaluation for tracking. Detection bounding boxes and ground truth person locations are provided for each sequence, with the ground truth data being used only for performance evaluation after running the tracker.

It should also be noted that our deep re-ID architecture is trained only on the MARS dataset [28], and not on images of people from either of the MOTC videos. Therefore, this experiment also studies the ability of our deep re-ID system to generalize to new images of previously unseen people.

For each sequence, several data association strategies are compared in a controlled study. The data association variants
we use are as follows: Using position only, using deep re-ID only, using a benchmark color histogram appearance model, and finally, using position along with deep re-ID. The benchmark appearance model uses a color histogram in HSV color space to judge visual similarity between tracked people and new detections, and is included in order to evaluate the performance afforded by deep re-ID as compared to a much simpler, non-machine learning-based model.

Performance is measured quantitatively with the widely used CLEAR MOT metrics, consisting of multiple object tracking accuracy (MOTA) and multiple object tracking precision (MOTP) [3]. MOTA decreases as the rate of false positives, false negatives, or ID switches increase. An example of an ID switch is shown in Figure 7, where the tracker mistakenly swaps the tracks of two different people. The hypothesis behind using deep re-ID for data association is that it should increase MOTA by decreasing the number of ID switches in particular, since ID switches are tracker errors that are directly caused by data association mistakes. In cases such as that seen in Figure 7, where two people walk next to one another, data association based on detection positions only is extremely ambiguous. A small amount of detection noise or unexpected motion could easily cause an ID switch. However, incorporating re-ID into data association should intuitively protect against such occurrences, as the tracker could then use appearance information to disambiguate nearby people from one another.

MOTP measures the precision with which people’s exact locations are known. This is primarily influenced by the precision of the object detector; a detector that fits bounding boxes more closely around people, or avoids bounding boxes entirely in favor of segmenting masks, would attain higher precision. On the other hand, the choice of data association strategy has no direct influence on precision, and so we do not expect a large effect on MOTP.

B. Quantitative Analysis

Tracking results from running the RBPF tracker on the two evaluation sequences are shown in Tables I and II. MOTA, MOTP, FP, FN, and ID Sw. indicate Multiple Object Tracking Accuracy, Multiple Object Tracking Precision, false positives, false negatives, and ID switches, respectively. The PETS09-S2L1 sequence includes 795 video frames showing 19 pedestrians walking, with a total of 4650 ground truth person annotations over all frames. This video sequence is fairly sparse; at any given time, 2 to 8 pedestrians are seen in the video frame simultaneously. The MOT17-04 sequence includes 1050 frames showing 83 pedestrians, with 47557 total ground truth annotations. This sequence is much more complex than the PETS09-S2L1 video; as seen in Figure 6, at times upwards of 30 people can be observed in the video frame, greatly increasing the difficulty of tracking.

The most significant effect of combining re-ID with

| Method          | MOTA | MOTP | FP   | FN   | ID Sw. |
|-----------------|------|------|------|------|--------|
| Pos. only       | 0.905| 0.644| 154  | 260  | 30     |
| Re-ID only      | 0.109| 0.544| 597  | 3434 | 112    |
| Pos.+Hist.      | -0.168| 0.560| 4801 | 411  | 218    |
| Pos.+re-ID      | 0.929| 0.656| 114  | 210  | 6      |

TABLE I
RESULTS ON SEQUENCE PETS09-S2L1

| Method          | MOTA | MOTP | FP   | FN   | ID Sw. |
|-----------------|------|------|------|------|--------|
| Pos. only       | 0.445| 0.820| 7831 | 18390| 160    |
| Re-ID only      | 0.156| 0.793| 2919 | 36684| 535    |
| Pos.+Hist.      | 0.195| 0.738| 21977| 15437| 862    |
| Pos.+Re-ID      | 0.533| 0.863| 2253 | 19871| 83     |

TABLE II
RESULTS ON SEQUENCE MOT17-04
position-based data association is a reduction in the number of ID switches, validating our earlier hypothesis. Compared to data association with position only, using position as well as re-ID caused a drop in ID switches from 30 to 6 in the PETS09-S2L1 sequence (an 80% reduction), and from 160 to 83 in the MOT17-04 sequence (a 48% reduction).

In order to realize the benefits of re-ID data association, it is not enough to use re-ID without position information, or to use a simple color histogram appearance model. Both approaches perform poorly. In the PETS09-S2L1 sequence, the color histogram method creates so many false tracks that its false positive count exceeds the total amount of ground truth annotations, causing negative MOTA overall. The poor performance of this method could be explained by the fact that color histograms are heavily influenced by colors in the background, as well as by the color of large objects such as a person’s coat. In contrast, a re-ID feature representation can be learned such that it ignores irrelevant information.

As expected, MOTP is not significantly affected by the inclusion of deep re-ID for data association as opposed to the position-only setting, though we do observe a modest increase. MOTP does however drop when MOTA is very low, as in the re-ID only and color histogram data association cases.

C. Qualitative Analysis

We next present a qualitative analysis of a scenario from the tracking results, in order to illustrate how deep re-ID assists data association and reduces the occurrence of ID switches, as seen from our quantitative results. Figure 8 shows a series of frames from the PETS09-S2L1 sequence, where a group of people walk across the scene. Tracking results with and without re-ID data association are shown, with the left-side column showing position-only data association tracking, and the right-side column showing tracking using position and re-ID.

In frame 150 of the position-only case, an ID switch occurs as one person (in the yellow bounding box) walks behind a signpost while another person, who previously had been lost by the tracker due to an extended period of occlusion, walks out from behind it at nearly the same location. The yellow box can be seen to have switched from one person to the other in frame 155. As the first person walks out from the other side of the signpost in frame 160, a second ID switch from a different person occurs, again due to occlusion, as seen from the movement of the purple box. In frame 165, the yellow box again switches between people, causing one person to be missed by the tracker in frame 170 until a new track is initialized in frame 175. When re-ID as well as position are used for data association, these ID switches do not occur, and the tracker is able to successfully use the distinct appearances of these people to perform reliable tracking even in the presence of occlusions and path crossings.

V. Conclusion

A general approach to augmenting traditional sensor likelihood models with deep person re-identification is presented, for application in multiple person tracking. We describe the process of converting images of people into convolutional feature vectors, using a learned deep re-ID model, and show how these feature vectors can be used within the general data association framework. Our results indicate that person re-ID significantly increases tracking performance as compared to data association that uses detection position only, according to quantitative measures of tracking accuracy and consistency. In particular, the usage of deep re-ID for data association is seen to cause an 80% drop in ID switches for tracking in the PETS09-S2L1 video sequence, as well as a 48% reduction in the much more crowded and complex MOT17-04 sequence. The usage of deep re-ID is additionally seen qualitatively to increase tracking robustness to difficulties such as occlusions and path crossings.
REFERENCES

[1] E. Ahmed, M. Jones, and T. K. Marks, “An Improved Deep Learning Architecture for Person Re-Identification,” CVPR, 2015.

[2] Y. Bar-shalom, F. Daum, and J. Huang, “The Probabilistic Data Association Filter: Estimation in the Presence of Measurement Origin Uncertainty,” IEEE Control Systems, vol. 29, no. 6, 2009.

[3] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The CLEAR MOT metrics,” EURASIP Journal on Image and Video Processing, 2008.

[4] S. S. Blackman, “Multiple Hypothesis Tracking,” IEEE Aerospace and Electronic Systems Magazine, 2004.

[5] D. Cheng, Y. Gong, S. Zhou, J. Wang, and N. Zheng, “Person re-identification by Multi-Channel Parts-Based CNN with Improved Triplet Loss Function,” CVPR, 2016. [Online]. Available: http://ieeexplore.ieee.org/document/7780518/

[6] J. Dai, Y. Li, K. He, and J. Sun, “R-FCN: Object Detection via Region-based Fully Convolutional Networks,” NIPS, 2016.

[7] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” CVPR, 2014.

[8] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN,” CVPR, 2017.

[9] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, “Densely Connected Convolutional Networks,” 2016. [Online]. Available: http://arxiv.org/abs/1608.06993

[10] M. Keuper, S. Tang, Y. Zhongjie, B. Andres, T. Brox, and B. Schiele, “A Multi-cut Formulation for Joint Segmentation and Tracking of Multiple Objects,” 2016. [Online]. Available: http://arxiv.org/abs/1607.06317

[11] J. M. Kim, Chanho; Li, Fuxin; Ciptadi, Arridhana; Rehg, “Multiple Hypothesis Tracking Revisited,” ICCV, 2015.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” NIPS, 2012.

[13] L. Leal-Taixe, A. Milan, I. Reid, S. Roth, and K. Schindler, “MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking,” 2015. [Online]. Available: http://arxiv.org/abs/1504.01942

[14] W. Li, R. Zhao, T. Xiao, X. B. T. C. V. Wang, and P. Recognition, “DeepReID: Deep Filter Pairing Neural Network for Person Re-identification,” CVPR, 2014.

[15] W. Li, X. Zhu, and S. Gong, “Person re-identification by deep joint learning of multi-loss classification,” IJCAI International Joint Conference on Artificial Intelligence, 2017.

[16] A. Milan, L. Leal-Taixe, I. Reid, S. Roth, and K. Schindler, “MOT16: A Benchmark for Multi-Object Tracking,” pp. 1–12, 2016. [Online]. Available: http://arxiv.org/abs/1603.00831

[17] A. Milan, S. H. Rezatofighi, A. Dick, I. Reid, and K. Schindler, “Online Multi-Target Tracking Using Recurrent Neural Networks,” AAAI, 2016. [Online]. Available: http://arxiv.org/abs/1604.05035

[18] I. Miller and M. Campbell, “Rao-Blackwellized Particle Filtering for Mapping Dynamic Environments,” Robotics and Automation, 2007 IEEE International Conference on, no. April, pp. 3862–3869, 2007.

[19] I. Miller, M. Campbell, and D. Huttonlocher, “Efficient unbiased tracking of multiple dynamic obstacles under large viewpoint changes,” IEEE Transactions on Robotics, vol. 27, no. 1, pp. 29–46, 2011.

[20] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137–1149, 2017.

[21] E. Ristani, F. Solera, R. Zou, R. Cucchiara, and C. Tomasi, “Performance measures and a data set for multi-target, multi-camera tracking,” ECCV, 2016.

[22] A. Sadeghian, A. Alahi, and S. Savarese, “Tracking The Untrackable: Learning To Track Multiple Cues with Long-Term Dependencies,” ICCV, 2017. [Online]. Available: http://arxiv.org/abs/1701.01909

[23] S. Sarkka, A. Vehtari, and J. Lampinen, “Rao-Blackwellized particle filter for multiple target tracking,” Information Fusion, vol. 8, no. 1, pp. 2–15, 2007. [Online]. Available: [http://resolver.scholarsportal.info/resolve/15662535/v08i0001/2]/rpfmlt

[24] D. Schulz, D. Fox, and J. Hightower, “People Tracking with Anonymous and ID-Sensors Using Rao-Blackwellised Particle Filters,” International Conference on Artificial Intelligence (IJCAI), pp. 921–926, 2003.

[25] Y. Wang, L. Wang, Y. You, X. Zou, V. Chen, S. Li, G. Huang, B. Hariharan, and K. Q. Weinberger, “Resource Aware Person Re-identification across Multiple Resolutions,” CVPR, 2018.

[26] M. Yang, Y. Wu, and Y. Jia, “A Hybrid Data Association Framework for Robust Online Multi-Object Tracking,” pp. 1–13, 2017. [Online]. Available: http://arxiv.org/abs/1703.10764

[27] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8, 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4587584

[28] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian, “Mars: A video benchmark for large-scale person re-identification,” ECCV, 2016.

[29] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, “Scalable Person Re-identification : A Benchmark Scalable Person Re-identification : A Benchmark,” ICCV, 2015.