Robust Global Tracker based on an Online Estimation of Tracklet Descriptor Reliability

Thi Lan Anh Nguyen, Duc Phu Chau, Francois Bremond

Robust Global Tracker based on an Online Estimation of Tracklet Descriptor Reliability

Nguyen Thi Lan Anh Chau Duc Phu Francois Bremond
INRIA Sophia Antipolis
2004 Route des Lucioles -BP93 06902 Sophia Antipolis Cedex - France
{thi-lan-anh.nguyen|duc-phu.chau|francois.bremond}@inria.fr

Abstract

The complex scene conditions such as light change, high density of mobile objects or object occlusion can cause object mis-detections. When a tracker can not recover these mis-detections, the trajectory of an object is fragmented into some short trajectories called tracklets. As a result, tracking quality is reduced remarkably. In this paper, we propose a new approach to improve the tracking quality by a global tracker which merges all tracklets belonging to an object in the whole video. Particularly, we compute descriptor reliability over time based on their discrimination. On the other hand, a motion model is also combined with appearance descriptors in a flexible way to improve the tracking quality. The proposed approach is evaluated on four benchmark datasets. The obtained results show the robustness and effectiveness of our approach compared to tracking as well as tracklet linking approaches from state of the art.

1. Introduction

Many recent approaches have been proposed to track mobile objects in a video. However, the quality of a tracker as well as the reliability of object descriptors strongly depend on video scenes. An object descriptor(*e.g*. color, size, motion...) can be useful in this scene but unreliable in other scenes. Therefore, in order to merge effectively tracklets, a dynamic combination of reliable object descriptors for tracking depending on video scenes is an important task.

In the state of the art, some online learning approaches have been proposed to track objects in various video scenes. For example, the authors in [5] present an online learning approach to adapt the object descriptors to the current background. However, the training phase requires user interaction which increases significantly the training cost. Besides, the approach in [4] tracks multi-objects by using tracklet confidence with an online discriminative appearance learning based on an incremental linear discriminant analysis (ILDA). This allows the tracker to learn discriminative appearance models and also incrementally update learned appearance models with tracking results.

However, the computation of object descriptors based on the current frame as above algorithms is less reliable than based on the trajectory in a number of frames. In order to deal with this issue, some recent researches focus on *global tracking* methods whose objective is to match short trajectories to create more completed object trajectories. The approach in [2] proposes an algorithm that recovers fragmentation of object trajectories by using enhanced covariance-based signatures and an online threshold learning. The authors in [13] propose a hierarchical relation hypergraph based tracker. In this approach, vertices are tracklets and edges are the relationship between tracklets. These global tracking algorithms have significant results in matching short trajectories. However, object descriptor weights are fixed for the whole video. Therefore, their tracking performances can be reduced if the scene change.

On the other hand, appearances of objects look similar to each other. Appearance based trackers in [9, 5, 4, 2] are less effective and less reliable while motion model based trackers in [13, 1] may become more useful. Therefore, an adaptive combination of appearance descriptors and motion model is necessary to improve the tracking quality.

In this paper, we propose a global tracker which computes and combines adaptively discriminative appearance descriptors and motion model to merge tracklets. The proposed approach brings three following contributions:

- A new model for describing a tracklet. In this model, each object descriptor is updated with object detections in many frames to increase its reliability.

- A method to compare and merge effectively tracklets by computing automatically discriminative descriptors depending on video scenes. Different from existing global tracking approaches, our approach does not fix the descriptor weights of tracklets for the whole video but computes and adapts them overtime.

- A flexible combination of appearance descriptors and

motion model to improve tracking quality.

The rest of this paper is organized as follows: Section 2 introduces a new model to describe a short trajectory. The proposed global tracking algorithm is presented in section 3. Section 4 shows evaluation results as well as analysis about the proposed approach performances. Finally, conclusions are summed up in section 5.

2. Tracklet model

In this section, we propose a new model for describing a tracklet. A reliable tracklet is determined as a short trajectory and should include following criteria (C) as well as information (I):

(I_1) **An identifier**:The tracklet ID.

(I_2)**Tracklet Descriptors:** Object descriptors are used to describe a tracklet. Tracklet descriptors are updated by a sequence of object detections to increase their reliability (more detail in section 2.2).

(I_3) **Overlapping tracklet list**: A list of tracklets having temporal-overlapping as well as spatial-overlapping with the given tracklet (more detail in section 2.3).

(I_4) **Matching candidate tracklet list**: A list of tracklets which could match with the given tracklet (more detail in section 2.4).

(C_1)**No Ghost Object**: A tracklet appearing in more than four frames is considered as reliable and is kept.

(C_2) **No Object Detection Anomaly**: A tracklet cannot have some object detections which change suddenly their appearances in several consecutive frames.

2.1. Tracklet Filtering

In order to ensure that tracklets are reliable, tracklets which do not satisfy the criteria C_1 and C_2 are filtered. With the criterion C_1, tracklets whose length is less than or equal to four frames are considered as noises and removed. With the criterion C_2, if object detections of a tracklet change their appearances in several consecutive frames, these object detections are removed and the given tracklet is fragmented.

2.2. Tracklet descriptors

Object descriptors play an important role to ensure a good tracking quality with different scene conditions. We propose to use a pool of the following descriptors to describe a tracklet.

2D shape ratio and 2D area Let W and H be the width and height of the 2D bounding box of an object. The 2D shape ratio, 2D area of this object are respectively defined as W/H and WH. If no occlusion occurs and objects are well detected, shape ratio and area of an object within a temporal window does not change much even if the lighting and contrast conditions are not good.

Color histogram and Dominant Color The color histogram of an object is defined as the normalized RGB color histogram of moving pixels inside its bounding box. Meanwhile, dominant color descriptor is similar to the color histogram descriptor, but it takes into account only the important colors of the object.

2D shape ratio, 2D area, color histogram, dominant color descriptors are updated by the means of the previous accumulative values and the values of the current object detection.

Motion descriptor Depending on the video context, we propose to use a *constant velocity model* or *Brownian model* from [7] to describe an object movement. In [7], motion model is represented by a Gaussian distribution. The motion model descriptor is useful for objects that have similar appearances.

Color covariance descriptor is defined in [12]. This is a very useful descriptor to characterize the appearance model of an image region. In particular, the covariance matrix enables to compare regions of different sizes and is invariant to identical shifting of color values. Therefore, this descriptor is even reliable when objects are tracked under varying illumination conditions. Since covariance matrices lay in a Riemannian manifold, we use the intrinsic Newton gradient descent algorithm to compute the approximate mean covariance over a time-window.

2.3. Overlapping tracklet list

Tracklet Tr^j is an overlapping tracklet of tracklet Tr^i if tracklet Tr^j has at least one frame overlapping with tracklet Tr^i (called as temporal-overlapping) and the 2D distance of both tracklets is below a predefined threshold (called as spatial-overlapping).

Overlapping tracklet list is figured out as a set of overlapping tracklets of tracklet Tr^i.

2.4. Matching candidate tracklet list

Tracklet Tr^p is determined as a matching candidate of tracklet Tr^i if tracklet Tr^p satisfies spatial-temporal constraints with Tr^i.

Suppose that Tr^i appears earlier than Tr^p. The *temporal constraint* ensures that the last object detection of Tr^i must appear earlier than the first object detection of Tr^p.

Spatial constraint ensures that the last object detection of Tr^i can reach the first object detection of Tr^p after a number of frames of potential mis-detection with the current frame rate.

Matching candidate tracklet list is determined as a set of matching candidates of tracklet Tr^i.

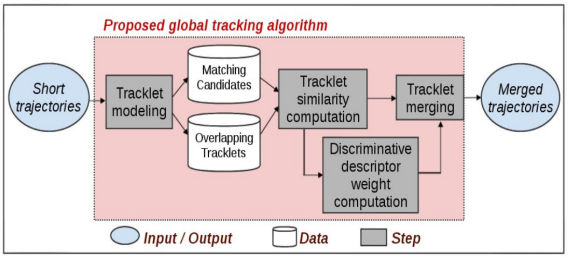

Figure 1. The overview of the proposed algorithm.

3. Proposed global tracking algorithm

The overview of the proposed algorithm is shown in figure 1. The algorithm takes the output of any multi-object tracker as input. With criteria C_1, C_2 of tracklet model, the input of the algorithm is reliable tracklets whose descriptors are updated over time. For each time-window Δt, the proposed approach considers all tracklets in a double sliding time-window $[t - 2\Delta t, t]$ and the global tracking algorithm is run once. Thanks to the overlap of one time-window, misdetections occurring inside a double time-windows can be recovered.

The objective of the proposed algorithm is to merge tracklets by using reliable descriptors which ensure that one tracklet can discriminate with its overlapping tracklets and still has a high matching reliability with its matching candidate. In this approach, the reliability of one tracklet descriptor is computed overtime and represented by a *discriminative descriptor weight*. Then, based on these weighted descriptors, the similarities between pairs of matching candidates (represented by the global matching scores) are computed. Finally, tracklets are merged with their best matching candidates after optimizing the global matching scores using Hungarian algorithm. Both of discriminative descriptor weights and matching scores are computed based on the tracklet descriptor similarity. Therefore, in the remaining part, we describe the algorithm by 3 following steps:

- Computing the descriptor similarity between tracklets.
- Computing online the discriminative descriptor weights.
- Merging tracklets.

3.1. Tracklet descriptor similarities (DS)

2D shape ratio (k=1) and 2D area (k=2) similarities

The similarities of 2D shape ratio and 2D area descriptors DS_k (k=$\{1,2\}$) between tracklet Tr^i and Tr^j are defined as follows:

$$DS_k(Tr^i, Tr^j) = min(D_k^i, D_k^j)/max(D_k^i, D_k^j) \quad (1)$$

where $D_k^i = W^i/H^i$ with k=1, otherwise $D_k^i = W^i H^i$; W and H are means of widths and heights of 2D bounding boxes.

Color Histogram (k=3) and Dominant Color (k=4) similarities

Because the dominant color descriptor is similar to the color histogram descriptor, we propose to use the Earth Mover Distance (EMD) to compute both color histogram and dominant color descriptor similarities.

Color covariance similarity (k=5)

In order to compare the color covariance descriptor of two tracklets, we use the distance measure proposed in [8] to measure the similarity of two corresponding covariance matrices.

Motion model similarity (k=6) We propose to use Kulback-Leibler divergence to compare between two Gaussian distributions which represent the motion model of two tracklets.

3.2. Online computation of discriminative descriptor weights

The discriminative descriptor weights of one tracklet must be directly proportional to the descriptor similarity of this tracklet with its matching candidate and inversely proportional to the descriptor similarity of the given tracklet with its overlapping tracklets. Given descriptor k (k=1..5), a tracklet Tr^i and a matching candidate Tr^p, we define a discriminative descriptor weight for this pair of tracklets as follows:

$$\omega_k^{i,p} = \alpha^{[DS_k(Tr^i, Tr^p) - \tilde{X}(DS_k(Tr^i, Tr^j)) - 1]} \quad (2)$$

where Tr^j is the overlapping tracklet of Tr^i. $\tilde{X}$ is the median of the similarities of tracklet Tr^i with its overlapping tracklets. The advantage of the median is that its value is not affected by a few of extremely big or small values. Therefore, the median is meaningful in coding the similarity of Tr^i with its overlapping tracklets even these similarity values are not distributed uniformly. Furthermore, in order to normalize a discriminative descriptor weight in the interval [0,1], we propose to map this weight to the function α^X where X = $[DS_k(Tr^i, Tr^p) - \tilde{X}(DS_k(Tr^i, Tr^j)) - 1]$. If $-2 \leq X \leq 0$, function α^X returns into [0,1]. We select α=10 in the experiment.

In our approach, the appearance descriptor and motion model weights are computed separately.

A combination of discriminative appearances and motion model descriptor

Different from appearance descriptors which follow criterion C_2, the motion of tracklets can change remarkably. So, we cannot use equation 2 to estimate the reliability of a tracklet motion descriptor.

Therefore, the approach proposes a new way to compute the motion model weight based on other appearance descriptors:

$$\omega_6^{i,p} = 0.5 - 0.5 \max_k (\omega_k^{i,p}) \quad k = 1..5 \quad (3)$$

By an inverse transformation in equation (3), we can combine appearance descriptors and motion model adapting to a variation of video scenes. If appearance descriptors are reliable enough to discriminate objects, the proposed approach takes into account appearance descriptors more importantly than the motion model. Inversely, when objects have similar appearance, the proposed tracker relies more on the motion model than appearance descriptors. However, the motion model is not too reliable as the object can change its direction frequently or measurement errors are caused by detection errors, calibration. Therefore, in order to use motion model effectively, in equation (3), the value of the motion model weight is set with a maximum value of 0.5.

3.3. Tracklet merging

The algorithm computes the global matching scores of a tracklet with its matching candidates. By computing and updating overtime discriminative descriptor weights, the tracker determines reliable descriptors in each video scene. The global matching score of tracklet Tr^i with each tracklet in its matching candidate list (represented by Tr^p) is summed up with the corresponding weight as follows:

$$MS(Tr^i, Tr^p) = \frac{\sum_{k=1}^{6} (\omega_k^{i,p} + \omega_k^{p,i}) DS_k(Tr^i, Tr^p)}{\sum_{k=1}^{6} (\omega_k^{i,p} + \omega_k^{p,i})} \quad (4)$$

After computing these global matching scores, we construct a matrix M = $\{m_{ik}\}$ with i=1..n, k=1..n, where n is the number of tracklets in current time interval [$t - 2\Delta$t,t]. $m_{ik} = MS(Tr^i, Tr^k)$ computed by equation (4) if tracklet Tr^k is in the candidate list of Tr^i. Otherwise, $m_{ik} = 0$. Then, Hungarian algorithm is used to optimize the tracklet merging process. However, the Hungarian algorithm can only find out the best match between 2 tracklets belonging to one moving object per time. The merging process fails if there are more than 2 tracklets belonging to an object. In order to overcome this limitation, the proposed approach applies Hungarian algorithm until there is no more possible matches. Therefore, by adjusting the size of time-window Δt, we can easily recover mis-detections over a long period of time.

4. Experimental results

We propose to use the output of the tracker in [6] as input because this tracker can be considered as a typical one when using a pool of object appearance descriptors to track objects. The proposed approach is tested on some sequences of four public datasets: PET2015, PETS2009, TUD_stadtmitte and TUD_crossing. The performance of this approach is compared with the tracker in [6], some other tracking and tracklet linking approaches from the state of the art.

We use CLEAR MOT metrics to evaluate the proposed method. The multiple object tracking precision (MOTP) evaluates the intersection area over the union area of bounding boxes. The multiple object tracking accuracy (MOTA) calculates the accuracy composed of false negatives, false positives and identify switching. In addition, some other metrics are proposed to use. Let GT be the number of trajectories in the ground-truth of the testing video. MT shows the ratio of mostly tracked trajectories, ML represents the ratio of mostly lost trajectories and PT is the ratio of partially tracked trajectories ($PT = GT - MT - ML$). We also propose to use the metric FG representing the number of *track fragments* to show the outstanding tracklet merging performance of our approach.

4.1. PETs datasets

With the dataset PETs2015, we choose the sequence $W1\_ARENA\_Tg\_TRK\_RGB\_1$ to test our approach because of its challenges. This sequence has 240 frames, there are only few people but their sizes are much changed, they have the variation of poses.

With the dataset PETS2009, we choose the sequence S2_L1, view 1 which has 794 frames containing 21 mobile objects with many occlusions and objects moving with different directions.

Figure 2 (six top images belong to PETs2009 dataset while three bottom images belong to PETs2015) illustrates the tracking performance related to the online computation of discriminative descriptor weights depending on each video scene. With the situation on three top images, tracklet ID3 (shown by yellow bounding box) and tracklet ID14(shown by red bounding box) are mis-detected because they cross each other at frame 140. Therefore, the proposed discriminative descriptor based tracking algorithm is applied to recover these mis-detections. The overlapped tracklets are visualized in black eclipses. Almost appearance descriptors of tracklets are similar but both objects move with opposite directions to each other. In this case, the proposed tracker recovers mis-detections thanks to the tracklet motion model with the weight value is nearly 0.4.

Three middle images show a different chunk of the PETs2009 sequence. Tracklet ID31 (described by yellow bounding box) and tracklet ID32 (described by light blue bounding box) move with similar trajectoriesbut their appearance colors are quite discriminative (by the color of hair and coat). The highest weight equals to 0.6 for dominant color and color histogram while the motion model weight is only 0.1. Therefore, the proposed tracker focuses mainly on dominant color and color histogram descriptors and is able to track objects correctly (see in frame 565).

Two objects in dataset PETs2015 also have the similar

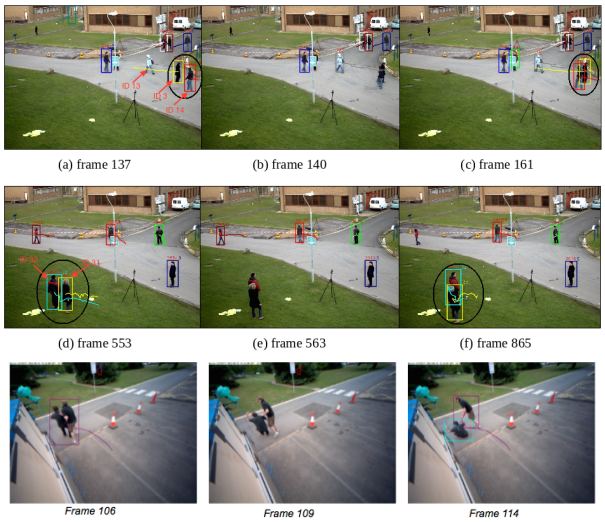

Figure 2. PETs2009 and PETs2015 datasets: The online computation of discriminative descriptor weights depending on each video scene.

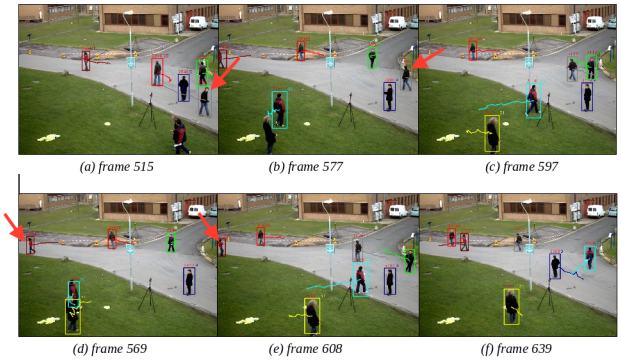

Figure 3. PETs dataset: Tracklet merging with the re-acquisition challenge.

appearance while having the different movement direction. In this case, the proposed approach relies mainly on object motion model to recover the trajectory fragmentation in frame 109.

Moreover, figure 3 shows our tracker's performance for the re-acquisition challenge when objects (shown by red arrows) leave and re-enter the scene. Instead of considering the moving objects in the frame they have just re-entered, our approach tracks these objects after a sufficient number of frames. Thanks to reliable descriptors which are updated cumulatively, object IDs are correctly retrieved.

4.2. TUD datasets

We also use TUD datasets (including TUD_Stadtmitte and TUD_crossing) sequences to evaluate the performance

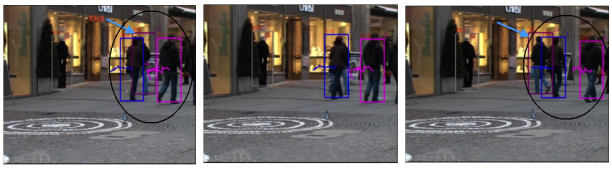

Figure 4. TUD-stadtmitte: The proposed approach performance in low light intensity condition, density of occlusion: Object ID_{26} (presented by purple bounding box) keeps its ID correctly after 11 frames of mis-detection.

of our approach compared to other recent trackers. Both of these sequences are quite short, with more or less than 200 frames, but they contain challenges for trackers such as low light intensity, crowded environment, frequent occlusions, similar object appearances.

Figure 4 illustrates clearly our approach performance when recovering mis-detection in videos that have low light intensity and objects moving in different directions. In these scenes, object appearances are not discriminative with each other. Appearance descriptors have similar discriminative weights around of 0.2 while the motion model weight is 0.4. Therefore, based on the motion model of objects, our approach tracks object ID26 (represented by a purple bounding box) correctly after several mis-detection frames.

The ***comparison*** is shown in table 1. With PETs2009 dataset, our performance is better than the tracker [4]-*with all its proposed methods* in MOTA and MOTP metrics and there is a significant tracking quality improvement when comparing our tracker with our input [6]. In particular, MOTA from 0.62 to 0.86 and 0.63 to 0.72 from MOTP, the track fragmentation (FG) reduce a half. Compared with other tracklet merging algorithms (in bold), our approach has a slightly lower results of MOTA and MOTP than [15] and [3]. However, the tracker in [15] works offline while our algorithm chooses flexibly object descriptors overtime which is suitable for the real-time applications. The tracker in [3] has a better performance than ours when it used its own input. When being tested with the same input (the output of tracker in [6]) with ours, our approach has higher results. Moreover, our performance is also much better than one of the tracker in [4] in case of the global association, especially in MOTA, MOTP and FG metrics.

On both the TUD datasets, our approach does not lose any object. The obtained ML values are also the best ones compared to other state of the art trackers in both datasets. Our tracker performance increases MT values from 60% to 70% with TUD_Stadtmitte and from 46.2% to 53.8% with TUD_Crossing dataset compared with our input.

The results show that the proposed global tracker is able to improve the tracking results by increasing the length and the precision of the tracklet. In particular, object trajecto-

Dataset	Method	MOTA	MOTP	GT	MT	PT	ML	FG
PETS2015	Chau *et al*. [6]	–	–	2	0.0	100.0	0.0	2
	Ours (Proposed approach + [6])	–	–	2	100.0	0.0	0.0	1
PETS2009	Chau *et al*. [6]	0.62	0.63	21	–	–	–	8
	Bae *et al.with all* [4]	0.83	0.69	23	100	0	0.0	4
	***Zamir* et al.** [15]	0.90	0.69	21	–	–	–	–
	***Bae* et al.**-*global association* [4]	0.73	0.69	23	100	0	0.0	12
	***Badie* et al.** [3]	0.90	0.74	21	–	–	–	–
	***Badie* et al.** [3] + [6]	0.85	0.71	21	66.6	23.9	9.5	6
	Ours (Proposed approach + [6])	0.86	0.72	21	76.2	14.3	9.5	4
TUD-Stadtmitte	Andriyenko *et al*. [1]	0.62	0.63	–	60.0	20.0	10.0	–
	Milan *et al*. [10]	0.71	0.65	9	70.0	20.0	0.0	–
	Yan *et al*. [14]	–	–	10	70.0	30.0	0.0	–
	Chau *et al*. [6]	0.45	0.62	10	60.0	40.0	0.0	13
	Ours (Proposed approach + [6])	0.47	0.65	10	70.0	30.0	0.0	7
TUD-Crossing	Tang *et al*. [11]	–	–	–	53.8	38.4	7.8	–
	Chau *et al*. [6]	0.69	0.65	11	46.2	53.8	0.0	14
	Ours (Proposed approach + [6])	0.72	0.67	11	53.8	46.2	0.0	8

Table 1. Tracking performance. The best values are printed in red, the second best values are printed in blue.

ries are more completed (Mostly Track (MT), MOTA and MOTP and FG values increase) while lost object trajectories are reduced (Mostly Lost (MT) values decrease).With FG metric, the proposed approach always has the least number of track fragmentation.

5. Conclusions

This paper proposes a new approach to improve the multi-object tracking quality by merging tracklets belonging to the same mobile object. In order to handle the variation of scenes, we propose a discrimination method to estimate the tracket descriptor reliability overtime. This ensures a high precision for tracklet merging process. Moreover, an adaptive combination of motion model and appearance descriptors is proposed to improve the tracker quality. The experimental results show the significant performance improvement of our approach compared to the input tracker, tracking as well as tracklet linking approaches from the state of the art over four experimented benchmark datasets.

Acknowledgement

This work is supported by the Provence-Alpes-Cote d'Azur Region, SafEE, Movement and PANORAMA projects.

References

[1] A. Andriyenko and K. Schindler. Multi-target tracking by continuous energy minimization, 2011. CVPR.

[2] J. Badie, S. Bak, S. Serban, and F. Bremond. Recovering people tracking errors using enhanced covariance-based signature, 2012. AVSS.

[3] J. Badie and F. Bremond. Global tracker: an online evaluation framework to improve tracking quality, 2014. AVSS.

[4] S. H. Bae and K.-J. Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning, 2014. CVPR.

[5] A. Borji, S. Frintrop, D. Sihite, and L. Itti. Adaptive object tracking by learning background context, 2012. CVPR.

[6] D. P. Chau, F. Bremond, and M. Thonnat. A multi-feature tracking algorithm enabling adaptation to context variations, 2011. ICDP.

[7] C. G. Cifuentes, M. Sturzel, F. Jurie, and G. J.Brostow. Motion models that only work sometimes, 2012. BMCV.

[8] W. Frstner and B. Moonen. A metric for covariance matrices, 1999. Quo vadis geodesia ...?, Festschrift for Erik W. Grafarend on the occasion of his 60th birthday.

[9] C. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by online learned dis- criminative appearance models, 2010. CVPR.

[10] A. Milan, S. Roth, and K. Schindler. Continuous energy minimization for multi-target tracking, 2014. PAMI.

[11] S. Tang, M. Andriluka, A. Milan, K. Schindler, S. Roth, and B. Schiele. Learning people detectors for tracking in crowded scenes, 2013. ICCV.

[12] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and clas- sification, 2006. ECCV.

[13] L. Wen, W. Li, J. Yan, Z. Lei, D. Yi, and S. Li. Multiple target tracking based on undirected hierarchical relation hypergraph, 2014. CVPR.

[14] X. Yan, I. A. Kakadiaris, and S. K. Shah. What do i see? modeling human visual perception for multi-person tracking, 2014. ECCV.

[15] A. R. Zamir, A. Dehghan, and M. Shah. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs, 2012. ECCV.