MULTI-TARGET TRACKING BY DISCRIMINATIVE ANALYSIS ON RIEMANNIAN MANIFOLD

Sławomir Bąk, Duc-Phu Chau, Julien Badie, Etienne Corvee, Francois Brémond, Monique Thonnat

INRIA Sophia Antipolis, STARS group
2004, route des Lucioles, BP93
06902 Sophia Antipolis Cedex - France
firstname.surname@inria.fr

ABSTRACT

This paper addresses the problem of multi-target tracking in crowded scenes from a single camera. We propose an algorithm for learning discriminative appearance models for different targets. These appearance models are based on covariance descriptor extracted from tracklets given by a short-term tracking algorithm. Short-term tracking relies on object descriptors tuned by a controller which copes with context variation over time. We link tracklets by using discriminative analysis on a Riemannian manifold. Our evaluation shows that by applying this discriminative analysis, we can reduce false alarms and identity switches, not only for tracking in a single camera but also for matching object appearances between non-overlapping cameras.

Index Terms— tracking, controller, re-identification, covariance matrix

1. INTRODUCTION AND RELATED WORK

Tracking multiple objects in real world scenes involves dealing with strong occlusions, illumination changes, and cluttered or moving backgrounds. In dense scenarios, similar object appearance and complex interactions between different targets often result in incorrect trajectories with fragmentation and identity switches.

Recent studies focus on *detection-based* tracking methods as the result of significant improvement in object detection algorithms. These tracking methods detect objects of interest, and then associate detection responses using several cues, such as appearance, motion, size, and other geometrical constraints. Unfortunately, missed detections and inaccurate responses frequently occur, which provides misleading information to tracking algorithms. Thus, *detection-based* tracking must overcome these difficulties by using different tracking strategies.

In [1], a two-stage approach (local and global) is presented. At the local stage, humans are tracked using body parts and particle filter. At the global stage, trajectories are associated by the *Hungarian algorithm*. The association cost matrix is computed using a simple appearance model based on color histogram, object height and velocity. A more complex appearance model is presented in [2]. This appearance model discriminates between the object and the background by employing Haar wavelet features and local binary patterns. Features are combined using a boosting scheme, in which negative samples are generated by randomly selected background regions.

Many approaches either focus on tracking strategies using simple features as color histograms, or develop the appearance models discriminating a target from the background [3]. However, few studies have been undertaken to resolve ambiguities between the different targets. In [4], the authors propose an online learning strategy which extracts appearance model for a set of targets in a sequence, to enhance the robustness of tracklet linking. This model is learned to maximize inter-tracklet variation whilst minimizing intra-tracklet variation.

For the last six years, the *person re-identification* problem has been the focus of intense research. Person re-identification can be seen as an extension of the tracking task to a multi camera scenario, in which the appearance of the same object registered in disjoint camera views has to be matched. Significant appearance changes, caused by variations in view angle, illumination and object pose, make the problem much more challenging than the single-camera tracking. As the person re-identification problem is particularly hard, the appearance models which tackle this problem should be able to handle difference in illumination, pose and camera parameters.

In this paper, we propose to take advantage of robust re-identification descriptor [5] for linking trajectories during the tracking procedure. We show that the re-identification descriptor which handles significant appearance changes across disjoint cameras, is able to merge tracklets extracted from a single camera.

This paper makes the following contributions:

- We propose to use *Mean Riemannian Covariance Grid* (MRCG) descriptor [5] for linking tracklets into longer

ones to form the final tracking results in a single camera. The tracklets are obtained by a controller-based short-term tracking algorithm (Section 3).

- We present a new approach for discriminative appearance learning based on a sliding time window. Each tracklet is learned to highlight its distinctive features. We show that this technique can be easily applied to tracking systems, without extracting any additional reference dataset for learning (Section 4).

The overview of our technique is given in Section 2. Section 3 describes a short-term tracking controller. Then, in Section 4 a discriminative learning is presented. We evaluate our approach in Section 5 before discussing future work and concluding.

2. OVERVIEW OF OUR APPROACH

The proposed approach consists of 3 main steps: (1) object detection; (2) short-term tracking and (3) linking trajectories using online discriminated appearances (see Figure 1).

The first step, object detection, can be achieved by applying simple motion detection algorithms or specialized object detectors *e.g*. a human detector. In our case we combine both: motion and HOG-based detector [6].

In the second stage, we generate tracklets associating detection responses by using a *short-term tracking* algorithm (Section 3). The short-term tracking is designed to produce reliable tracklets, by employing spatial, temporal and appearance descriptors. We develop a *controller* [7], which assigns specific weights to these descriptors depending on the context of the video, thus generating reliable tracklets.

Finally, the last step of our approach consists in linking the tracklets using MRCG descriptor. We propose a linking strategy examining similarity between tracklets in a sliding time window. For each tracklet computed in sliding time window w, we build its MRCG representation. MRCG was designed to discriminate one appearance *vs* the reference appearances. We propose to discriminate each tracklet *vs* the rest of tracklets computed in given w. At that stage, the distinctive features of the tracked objects are enhanced, thus improving the linking accuracy. Unlike [4], for each tracklet we learn online the specific representation of the tracklet appearance, highlighting distinctive features of the particular object. It means that for N tracklets, we obtain N models, each highlighting its differences. In [4] only one model for a set of targets is learned to maximize discriminative appearance matching.

3. SHORT-TERM TRACKING

The short-term tracking is supervised by a *controller* [7] which automatically tunes the object descriptors to cope with the scene context variations.

3.1. Tracking Algorithm

The tracking algorithm takes as input the video stream and a list of objects detected in a sliding time window. First, a link score is computed between any two detected objects appearing in this time window using a linear combination of 8 object descriptors $\mathcal{T}$ extracted from 2D, 3D geometrical information, various colour descriptors and gradient cues (*i.e*. HOG). Successive links form several paths on which an object can undergo within this temporal window. Each possible path of an object is associated with a score given by all the scores of the links it contains. The object trajectory is simply determined by maximizing the path score.

3.2. Controller

The aim of the controller is to automatically adapt the object descriptors $\mathcal{T}$ to the context of a video. We train our controller with various scene contexts and then we employ the controler to improve tracking accuracy.

Training Phase. The training phase generates the context clusters. Videos are divided into video chunks where the context of each chunk is assumed constant. Using manually annotated objects (ground truth), video chunk contexts are modeled by 6 contextual features: the spatial density of detected objects; the object occlusion level; the object contrast with regard to the surrounding background; the variance of this contrast; the object $2D$ area and the variance of this area. The Quality Threshold (QT) clustering algorithm is used to cluster all the video chunk clusters from all the training videos. The Adaboost algorithm is then used for computing weights of each object descriptor for each context cluster to maximize object tracking performance. In the result, the training algorithm produces a learned database which contains several video context clusters associated with the best object descriptor weights.

Online Control Process. The goal of the online control process is to tune the weights of object descriptors $\mathcal{T}$ over time using the learned data from the offline training phase. For every video chunk of a predefined number of frames, we determine the context cluster membership. Depending on this cluster membership, we select the learned weights of object descriptors $\mathcal{T}$ to guide the tracking process. In the result, the tracking algorithm is able to cope with the context variations, thus producing reliable tracklets.

4. ONLINE DISCRIMINATED APPEARANCES

4.1. Appearance model

Short-term tracking algorithm generates short but reliable tracklets. A tracklet is represented by a set of cropped images corresponding to the tracked regions. We use this set of images for computing MRCG descriptor.

MRCG [5] refers to *Mean Riemannian Covariance Grid*, which has been designed to describe a set of images. MRCG

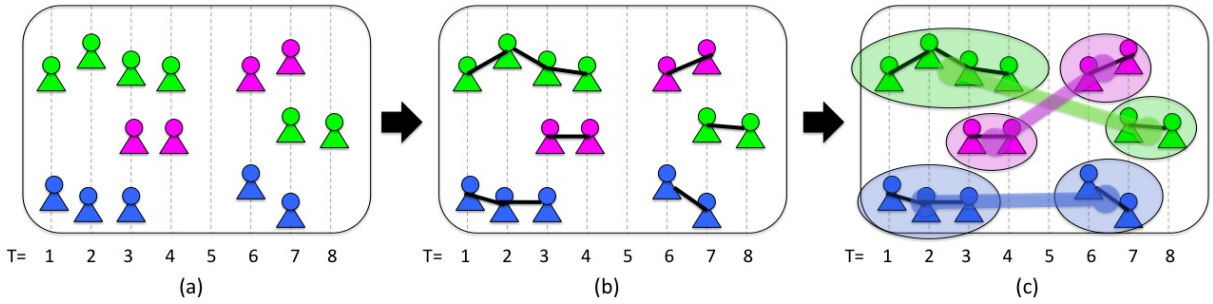

Fig. 1. The overview of the tracking approach in T frames: (a) The raw detection results; (b) the results of the short-term tracking; (c) the results of linked trajectories using online discriminated appearances.

forms a dense grid structure with spatially overlapping square regions (*cells*). These *cells* are described using mean covariance matrix. Since covariance matrices do not form a vector space, this mean covariance is computed on a Riemannian manifold. The mean covariance is an intrinsic average, which blends appearances from multiple images, holding information on feature distribution, their spatial correlations and their temporal changes during tracking. In the result, each tracklet is represented by MRCG. We employ a discriminative method to enhance distinctive characteristics of each tracklet w.r.t. the others, thus improving linking accuracy.

4.2. Discriminative method

The discriminative method assigns weights to MRCG *cells* of a specific object, reflecting their relevance. Given a set of signatures $\mathfrak{S}^w = \{\mathfrak{s}_i\}_{i=1}^n$, where s_i is signature i computed from the tracklet registered in sliding time window w, we represent MRCG by $\mathfrak{s}_i = \{\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{i,m}\}$, $\mu_{i,j}$ is the *mean Riemannian covariance* and m is the number of *cells* in the grid. For each $\mu_{i,j}$ we compute its relevance defined as variance $\sigma_{i,j}$ in a set of tracklet signatures extracted in window w. This variance is computed on a Riemannian manifold, projecting $\mu_{k,j}$ on tangent plane at $\mu_{i,j}$

$$\sigma_{i,j} = \frac{1}{n-1} \sum_{k=1;k\neq i}^{n} \| \log_{\mu_{i,j}}(\mu_{k,j}) \|^2_{\mu_{i,j}}, \qquad (1)$$

where operator $log_{\mu_{i,j}}$ is uniquely defined on the Riemannian manifold, enabling distance computation. Using this discriminative technique, we focus on the features which are distinctive and at the same time we disregard common patterns (often corresponding to background). Each signature is discriminated *vs* the rest of signatures computed in sliding time window w.

4.3. Linking the tracklets

Tracklet Similarity: Appearance of each tracklet is represented by MRCG signature. The similarity between two tracklet signatures $\mathfrak{s}_A$ and $\mathfrak{s}_B$ is defined as

$$S(\mathfrak{s}_A, \mathfrak{s}_B) = \frac{1}{|K|} \sum_{i \in K} \frac{\sigma_{A,i} + \sigma_{B,i}}{\rho(\mu_{A,i}, \mu_{B,i})} \qquad (2)$$

where K stands for the set of *cells* in the grid structure; ρ is the geodesic distance (see [5] for details); $\sigma_{A,i}$ and $\sigma_{B,i}$ are the weights of the corresponding *cells* computed by discriminative method (Section 4.2).

Linking: Given a set of tracklets, we search for the possible matching candidates based on temporal constraints. We assume that two tracklets can not overlap in time to be a candidate for linking. This assumption is based on the observation that one object can not belong to two different trajectories at the same time. Using *tracklet similarity* we link trajectories based on the threshold strategy applied to our similarity function. Two tracklets are linked together when their similarity is high enough ($S(\mathfrak{s}_A, \mathfrak{s}_B) > \theta$). Threshold θ is learned by maximizing true matches, while using the data employed for *controller* training.

5. EXPERIMENTAL RESULTS

We evaluate the effectiveness of our approach using two public surveillance datasets: CAVIAR[1] dataset and 2008 i-LIDS Multiple-Camera Tracking Scenario (MCTS)[2] dataset.

5.1. Single camera tracking - CAVIAR data

Caviar dataset contains 26 videos, 6 of them are used for training our controller and the remaining 20 are used for evaluation. We carry out an experiment employing the commonly used metrics [4]: GT - the number of trajectories in the ground truth; MT - the percentage of trajectories that are successfully tracked for more than 80% divided by GT; PT - the percentage of trajectories that are tracked between 20% and 80% divided by GT; ML -the percentage of trajectories that are tracked for

[1]CAVIAR: http://homepages.inf.ed.ac.uk/rbf/CAVIARDATA1
[2]i-LIDS: http://www.homeoffice.gov.uk/science-research/hosdb/i-lids/

Method	GT	MT (%)	PT (%)	ML (%)
Wu *et al.*[8]	140	75.7	17.9	6.4
Xing *et al.*[9]	140	84.3	12.1	3.6
Huang *et al.*[10]	143	78.3	14.7	7.0
Li *et al.*[11]	143	84.6	14.0	1.4
Kuo *et al.*[4]	143	84.6	14.7	0.7
Our approach	140	**84.6**	9.5	5.9

Table 1. Tracking results on Caviar dataset.

less than 20% divided by GT. The tracking results are presented in Table 1. Our results show that, although detection responses are missed (high ML), our discriminative method achieves high MT = 84.6%, reaching state of the art performances.

5.2. Multi camera tracking - i-LIDS data

We perform two experiments on i-LIDS data with multi cameras. The evaluation is presented in the light of linking the tracklets across disjoint camera views.

i-LIDS-AA [12]: This dataset contains 100 individuals registered in two non-overlapping cameras. For each individual a different number of cropped images is given, forming the tracklet. Our aim is to link correctly the tracklets from the first camera with the tracklets from the second camera. Although i-LIDS-AA was originally extracted for evaluating the *person re-identification* problem, this dataset can also be applied to test our approach. In experiments we set sliding window $w = 10$ individuals (based on the order in the dataset), which simulates our sliding time window. Assuming a regular flow of individuals from the first camera to the second camera, we successfully linked 72% of tracklets. It is worth noting that the best performance for re-identification achieved on this data reached 43% for the first rank in CMC curve [5]. The results show that our discriminative learning can handle such challenging aspects as different color responses and different camera settings.

i-LIDS-crowded:[3] This dataset contains a dense scenario with strong occlusions and complex interactions between objects. Crowded environment makes the object detection and the object tracking very challenging. We applied our short-term tracking to obtain the tracklets from both cameras. During linking the tracklets from the first camera with the tracklets from the second camera we tuned the similarity threshold to ensure 100% precision. Finally, 33.3% of ground-truth objects were linked together across disjoint camera views.

6. CONCLUSION

We proposed a new approach for linking tracklets in a single and multi camera scenario. By applying discriminative learning on appearances registered in a sliding time window, we are able to enhance the linking accuracy. The approach was evaluated on CAVIAR and i-LIDS datasets. In the future we will investigate how to improve object detection responses to minimize ML metric.

[3]TrecVid/Dev08/: 2007-11-01-CAM1-2 and 2007-11-01-CAM3-2

Acknowledgements

This work has been supported by VANAHEIM and VICOMO project.

References

[1] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *CVPR*, 2009.

[2] L. Snidaro, I. Visentini, and G.L. Foresti, "Dynamic models for people detection and tracking," in *AVSS*, 2008.

[3] R.T. Collins, Y. Liu, and M. Leordeanu, "Online selection of discriminative tracking features," *TPAMI*, 2005.

[4] Ch. Kuo, Ch. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *CVPR*, 2010.

[5] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid," in *AVSS*, 2011.

[6] E. Corvee and F. Bremond, "Body parts detection for people tracking using trees of histogram of oriented gradient descriptors," in *AVSS*, 2010.

[7] D. P. Chau, "Dynamic and robust object tracking for activity recognition," in *PhD thesis*, march 2012.

[8] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *IJCV*, 2007.

[9] J. Xing, H. Ai, and S. Lao, "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses," in *CVPR*, 2009.

[10] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," in *ECCV*, 2008.

[11] Y. Li, C. Huang, and R. Nevatia, "Learning to associate: Hybridboosted multi-target tracker for crowded scene," in *CVPR*, 2009.

[12] S. Bak, E. Corvee, F. Bremond, and M. Thonnat, "Boosted human re-identification using riemannian manifolds," *IMAVIS*, 2011.