Recovering people tracking errors using enhanced covariance-based signatures

J. Badie, S. Bak, S.T. Serban and F. Brémond
INRIA Sophia Antipolis, STARS group
2004, route des Lucioles, BP93 06902 Sophia Antipolis Cedex - France
{julien.badie | slawomir.bak | silviu-tudor.serban | francois.bremond}@inria.fr

Abstract

This paper presents a new approach for tracking multiple persons in a single camera. This approach focuses on recovering tracked individuals that have been lost and are detected again, after being miss-detected (e.g. occluded) or after leaving the scene and coming back. In order to correct tracking errors, a multi-cameras re-identification method is adapted, with a real-time constraint. The proposed approach uses a highly discriminative human signature based on covariance matrix, improved using background subtraction, and a people detection confidence. The problem of linking several tracklets belonging to the same individual is also handled as a ranking problem using a learned parameter. The objective is to create clusters of tracklets describing the same individual. The evaluation is performed on PETS2009 dataset showing promising results.

1 Introduction

Multiple persons tracking is an important and challenging task in video surveillance and one of its extensions is re-identification. The objective of re-identification consists in linking objects throughout a network of overlapping or non-overlapping cameras. One of the other possible extensions is to track with the same ID individuals that reappear after a long occlusion or after reentering the scene. This extension can be called re-acquisition or global tracking.

Global tracking can perform two different roles depending on the time window where it is applied. On the first hand, it can be used to correct tracking errors such as changing IDs due to long-term occlusions in a short-term point of view. On the other hand, it can be used to to track people under the same ID even if they appear several times on the same scene, for example during a day (figure 1). The proposed approach tries to answer to both of the problems with the same framework.

Considering the problem of global tracking, it is relevant to use a descriptor which is independent not only from the person's posture, position or activity but also from the characteristics of the scene, such as changing background. Therefore, appearance model based descriptors seem suited to identify people in video surveillance. Another constraint is that each tracklet (segment of trajectory) should be represented by a discriminative enough signature. This signature (representation of the appearance) helps to link several tracklets belonging to the same person.

In the state of the art, many different appearance-based descriptors with satisfying results already exist. A global tracking approach is described in [6] to correct lost trajectories thanks to learned scene semantic information. In [14], a tracking method based on body parts is proposed using edgelet features. In [12], Kuo *et al.* present a reliable descriptor and tracklet association method. However, in the majority of cases, two problems remains unsolved : the discrimination of the visual signatures (except in [12]) and the size of the time window where the algorithm can fuse two trajectories. To summarize, the state of the art has focused more on repairing trajectory interruptions due to short-term occlusions than checking if a person has left and reentered the scene.

Mean Riemannian Covariance Grid (MRCG) descriptor [2] is used in the case of re-identification on overlapping or non-overlapping cameras promising results outperforming the state of the art. However, this approach has two main limitations. The first limitation is that background pixels are also used to compute tracklet signatures. The second one is that the results are given as a list of the best match for one tracklet and not as a definitive link between two tracklets. We propose to adapt this descriptor in a single camera in a real-time situation while overcoming the limitations.

This paper makes the following contributions:

- We propose to use *Mean Riemannian Covariance Grid* (MRCG) descriptor for short-term error recovering and for long-term re-acquisition by linking several tracklets belonging to the same person in a real-time situation (Section 2.3) on a mono-camera system.
- We enhance the quality and the reliability of the signature using an adaptive background subtraction and a people detection confidence (Section 2.1 and 2.2).
- We propose a new method for linking person tracklets

(a) frame 194 (b) frame 213 (c) frame 233

Figure 1: The global tracking challenge : correcting errors due to occlusions (ID 142 on the first frame becomes 147 on the last frame) and track people that are levaing the scene and reentering (ID 133 on the first frame becomes 151 on the last frame)

and for creating clusters of tracklets using an adaptive parameter that can be learned in a real-time situation (Section 2.5).

The rest of the paper is organized as follows : Section 2 presents an overview of the proposed system, the features used to compute the tracklet signatures and how the tracklets are compared, linked and clustered. In Section 3, we discuss the results of the proposed method and compare them with the state of the art approaches. Section 4 concludes and presents some future works.

2 Proposed approach

The architecture of the proposed system is presented in Fig. 2. It is composed of the following steps : (1) object segmentation; (2) people detection; (3) short-term tracking; (4) global tracking. The global tracking gathers the outputs from all the previous algorithms and combine them to create the signature database of all non-noisy tracklets while the short-term tracker provides short but reliable tracklets in a small time window.

This architecture is flexible enough to adapt itself to any kind of short-term tracker. The global tracker post-processes the results of the short-term tracker by computing the tracklets signatures and choose to link (or not) the trackets.

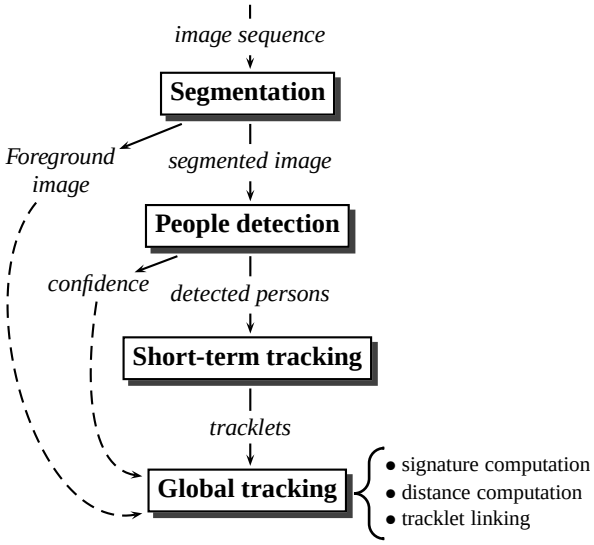

Figure 2: System architecture

2.1 Foreground image

We use a segmentation method based on an Extended Gaussian Mixture Model (EGMM) [13]. This method consists in detecting foreground pixels which are post-processed to remove shadows and highlights and adding a controller to adapt the background subtraction to the current scene condition. This method constructs the background representa-

tion with two features : chromaticity and brightness. This representation is updated every frame.

The background subtraction images are used as a mask to compute the foreground images (Fig. 3) and is used as an input of the global tracker.

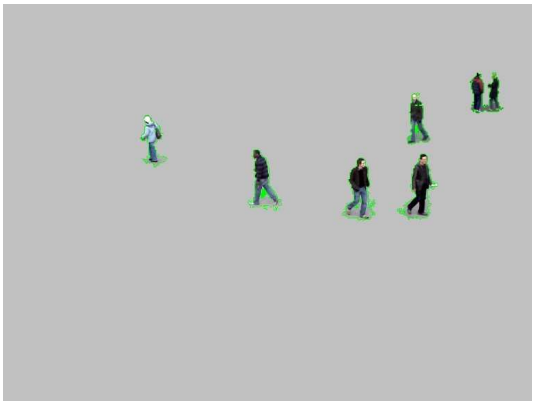

Figure 3: Foreground image with in green the contours

2.2 People detection confidence

In some cases, foreground images do not give a satisfying result, considering they will be used to compute a visual signature. For example, if on some frames, a person is partially or totally occluded, these frames should not be used to compute the person's visual signature. In order to handle this problem, the result of the segmentation are filtered by a specific people detection method.

For people detection, we propose to generate a value representing the confidence level of each detected object to measure the accuracy of the detection. In the proposed approach, we use a LBP-based method [10] to obtain several detection candidates (Fig. 4a) for one person. A matching score S is associated to each candidate. By adding a Mean Shift clustering algorithm to successfully group the candidates, we obtain a single bounding box representing the detected object.

Each matching score S is generated using Adaboost with a learned database. In theory, the value of S can be between 0 and 1. However, only the best candidates ($S > 0.5$) are selected and used to compute the confidence. Considering the average matching score S_a of all the N candidates and the best matching score S_b of all the candidates, the confidence C is computed using the following formula :

$$C = 1 - \frac{1 - (S_a + S_b)/2}{N} \quad (1)$$

This equation combines the LBP scores for human detection giving a hightest priority to the best match. The higher the confidence, the more correct the person detection (Fig. 4).

(a) (b) $C = 0.95$ (c) $C = 0.85$ (d) $C = 0.77$

Figure 4: Confidence computation : (a) set of candidates for one person; (b) : a perfectly detected individual with a high confidence; (c) : an uncertain detection with an average confidence ; (d) : a bad detection with a low confidence due to occlusion.

The drawback of the confidence is that it is not affected if more than one person is detected at the same place. As a consequence, it cannot be used to process occlusions between several individuals because it can only guarantee that at least one person is detected.

2.3 Signature computation

The short-term tracking algorithm is performed with a multi-feature tracker [7] using 3D position, shape, dominant color and HOG descriptors. This tracker generates short but reliable tracklets. A tracklet is represented by a set of cropped images corresponding to the tracked regions. This set of images is used to compute MRCG signature. In the proposed approach, the short-term tracker is only consider as a tracking reference whose results need to be improved.

MRCG descriptor [2] is used for re-identification in overlapping or non-overlapping cameras and shows very promising results. However, the reliability of the descriptor significantly depends on the quality of the input set of images. If the detection is not correct, the quality of the computed signature can significantly drop. In the original approach, the whole bounding box of the person is used to compute the signature (Fig. 5a). In our approach, we propose to combine background subtraction as a mask to keep only foreground pixels (Fig. 5b) and the people detection confidence to eliminate inaccurate detections with a confidence below a given threshold.

By filtering the images this way, we ensure that only significant pixels are used to compute the signature and we limit the number of noisy images. However, the quality of the signature is still limited by the quality of the background subtraction algorithm and the accuracy of the people detector.

Using MRCG descriptor online in single camera requires to save several images in every frames. We propose a method working as follows : we only consider non-noisy

(a) (b)

Figure 5: Input images to compute MRCG signature. (a) : original approach; (b) : our approach

trackets (with at least 5 frames). When the tracklet is considered as finished by the short-term tracker, we compute its signature using the foreground images corresponding to a reliable enough detection (C superior to a predefined threshold). Considering the current signature, we compute its distance with all the previously computed signatures using the distance defined in [2]. At any time, the global tracker has access to the current signatures and to the signatures that were lost.

Inspired by the method described in [12], distances (d) between all the combination of signatures are computed and arranged into two categories : possible links pl and impossible links il. Two tracklets are considered impossible to be linked if they are overlapping on at least one frame. For each category, we sort the links by increasing distance to obtain a tracklet similarity ranking.

2.4 Tracklets linking

The list of possible links $PL = \{pl_k\}_{k=1}^{N}$ and the list of impossible links $IL = \{il_k\}_{k=1}^{M}$ contains all the distances d between every 2-combination of trackets. Only the best possible links are used as candidates to be linked. To define these best possible links, we consider the highest rank of similarity between two tracklets that are impossible to link. This number is the smallest distance in the list of impossible links IL and is used as an initial threshold : $T_1 = d(il_1)$.

Considering this threshold, The list of possible links between tracklets is refined as follows :

$$PL' = \{pl_k : d(pl_k) < T_1\}_{k=1}^{N'} \tag{2}$$

As defined before, the impossible link list is established online and based on a temporal constraint. The threshold T_1 is updated with a new value if new impossible links appear and therefore, all previous links are updated using the new threshold. This threshold represents the adaptive parameter of the proposed approach.

The way the threshold is defined allows us to consider that the possible links list PL' has a limited number of wrong links. However, we may also ignore some correct links. To increase the sensitivity of the tracklet linking algorithm, a new threshold T_q is defined as follows :

$$T_q = \frac{1}{q}\sum_{k=1}^{q} d(il_k) \tag{3}$$

The parameter q is set manually and represents the number of impossible links that are considered to compute the threshold. As parameter q increases, the number of correct links also increases but we have also a greater chance to accept wrong links. The threshold T_q is only based on the impossible link list. Its reliability depends on the number of people that are detected in the scene at the same time and their similarity. Considering that it can be inappropriate to some sequences, it is also possible to learn the threshold in an offline phase using a database of images of different people.

Starting with the first rank PL', we use a Mean Shift algorithm to create clusters of tracklets that should correspond to the same person. Considering an existing cluster of n tracklets $[\tau_1, \cdots, \tau_n]$, and a possible link $[\tau_a, \tau_b]$, the linking condition is :

$$\frac{1}{2n}\sum_{k=1}^{n} d([\tau_a, \tau_k]) + d([\tau_b, \tau_k]) < T_q \tag{4}$$

If the linking condition is true, the possible link $[\tau_a, \tau_b]$ is added to the cluster. If the linking condition is wrong for all clusters and the tracklets τ_a or τ_b do not appear in any existing cluster, a new cluster is created with these two tracklets. Figure 6 shows an example of the linking algorithm inputs and outputs.

3 Experimental Results

We evaluate the effectiveness of the proposed approach using the public dataset PETS2009, on the particular sequence S2_L1, composed of 7 overlapping cameras recording 12 different people walking. This dataset is particularly relevant for the proposed approach because it contains a lot of occlusions, people that leave the scene and come back later and a number of people high enough to perform the linking algorithm without leaning the adaptive parameter with an offline leaning phase.

However, the default ground truth [3] shows 21 trajectories, using two different IDs to describe a person leaving and reentering in the scene. For performance evaluation, two different ground-truth data are used, a first one with 21 trajectories and a custom ground-truth with 12 trajectories taking into account the fact that people can reenter in the scene.

In order to compare the enhanced signatures with the original one, we evaluate the percentage of tracklets that

Figure 6: Linking algorithm inputs and output. The length of the output tracklet is the sum of the lengths of the inputs tracklets. In this example, the input tracklets length is between 8 and 120 frames and the output length is 348 frames.

are correctly linked, incorrectly linked and not linked according to the parameter q of the equation 3. A tracklet is considered as correctly linked if it is classified in a cluster representing the same person. An incorrect link occurs when the tracklet is put in a cluster representing a different person. Finally, not linked tracklets correspond to tracklets that are not assigned to any cluster whereas they should be. Our first results in Table 1 show the effectiveness of our approach on the first camera view (View_001), in the case when the 21 ground-truth trajectories are used as tracklets. Table 2 shows the results when the tracklets are computed with the short-term tracker.

The results show that the tracklets with an enhanced signatures are more likely to be added to a correct cluster (78.9%) compared to the state of the art signatures (59.7%). The proposed approach also decreases the error rate from 12.4% to 6.5%.

Some specific tracking metrics are presented in [5] and [11] for PETS2009 dataset. The computation of these metrics is reported in table 3. The value of the MOTP metric is low because the tracker only focus on creating reliable trackets and will fail to create this kind of tracklets during an occlusion. However the metrics *Multiple Object Tracking Accuracy* (MOTA) and *Multiple Object Tracking Precision* (MOTP) are not really adapted with the proposed method. These metrics works as follows : if the same person is described with two different ID, it is counted as one single error, not taking into account the length of both tracklets. Considering all the other possible errors (miss-detection, tracking errors), the influence of one ID switch errors does not appear clearly on these metrics. As a matter of fact,

Metrics	MODA	MODP	MOTA	MOTP
Berclaz *et al.*[4]	0.84	0.53	0.82	0.52
Yang *et al.*[15]	0.759	0.544	0.76	0.538
Conte *et al.*[9]	0.833	0.645	0.830	0.638
short-term tracker	0.8274	0.571	0.8271	0.327

Table 3: PETS metrics from [11] for the tracker alone

the values of these metrics are not significantly influenced enough by the proposed error recovering method. One solution to improve these metrics would have been to reconstruct the trajectory during the interval between two IDs of the same person using a trajectory optimization algorithm.

In order to evaluate successfully the performance of the proposed method, we use other evaluation metrics described in [14]. These metrics, implemented in the evaluation framework ViSEvAL [1], rely on :

- Mostly Tracked trajectories (MT) when more than 70% of the trajectory is tracked
- Partially Tracked trajectories (PT) when between 20% and 70% of the trajectory is tracked
- Mostly Lost trajectories (ML) when less than 20% the trajectory is tracked)

Table 4 shows the results using the original ground-truth including 21 trajectories. In this case, only occlusions or

Method	Tracklets	q	Correctly linked	Incorrectly linked	Not linked
MRCG [2]	21	1	12.5%	0%	87.5%
	21	5	50%	0%	50%
MRCG + foreground image	21	1	62.5%	0%	37.5%
	21	5	87.5%	0%	12.5%

Table 1: Global tracker re-acquisition rate based on ground truth tracklets of PETS2009 S2.L1.. Ground-truth tracklets are considered as perfectly reliable ($C = 1$)

Method	Tracklets	q	Correctly linked	Incorrectly linked	Not linked
MRCG [2]	129	1	21.7%	1.6%	76.7%
	129	5	53.5%	3.9%	42.6%
	129	10	59.7%	12.4%	27.9%
MRCG + confidence	76	1	23.3%	0.8%	75.9%
	76	5	55%	2.3%	42.7%
	76	10	60.5%	11.6%	28.3%
MRCG + confidence + foreground image	76	1	51.4%	0%	48.6%
	76	5	71.1%	5.2%	23.7%
	76	10	78.9%	6.5%	14.6%

Table 2: Global tracker re-acquisition rate based on the short-term tracker for the sequence S2.L1. from PETS2009. The number of tracklets corresponds to how many tracklets are used to compute the signatures. In the proposed approach, the confidence is used to filter the noisy tracklets.

Method	q	MT	PT	ML
D. P. Chau *et al.* [8]	–	14.3%	57.1%	28.6%
short-term tracker	–	9.5%	57.1%	33.3%
short-term tracker + global tracker	1	19%	61.9%	19%
	10	23.8%	57.1%	19%

Table 4: Tracking performance on PETS2009 S2.L1 View_001 sequence using the original ground-truth with 21 trajectories

Method	q	MT	PT	ML
D. P. Chau *et al.* [8]	–	8.3%	58.3%	33.3%
short-term tracker	–	0%	41.7%	58.3%
short-term tracker + global tracker	1	33.3%	41.7%	25%
	10	50%	33.3%	16.7%

Table 5: Tracking performance on PETS2009 S2.L1 View_001 sequence using the custom ground-truth with 12 trajectories

miss-detections can interrupt a tracklet. Although the short-term tracker used is not better than a state of the art tracker based on OpenCV Kalman filter [8], the global tracker sightly improves the results of the short-term tracker. The results shows that some tracklets can be merged after an occlusion. The tracklets of the Kalman filter based tracker are not used as an input for the global tracker because they are not reliable enough compared to the ones provided by the short-term tracker. Table 5 shows the results using the custom ground-truth including 12 trajectories. In this case, people leaving and reentering the scene are considered as the same person. It is normal that trackers [7] and [8] have only a small percentage of MT (8.3% and 0%). However, the proposed global tracker significantly improves this percentage up to 33.3% when $q = 1$ and up to 50% when $q = 10$.

4 Conclusions

In this paper we have presented a new approach for recovering errors using a global tracking method based on an enhanced appearance signature and a new strategy for linking tracklets. It has been shown that by using background subtraction and people detection confidence, we significantly improve the quality and the reliability of the results on PETS2009 dataset, while keeping a low level of error. Transforming the initial ranking problem into a decision problem is also performed using an adaptive parame-

ter. However, even if the proposed approach is independent from the tracker, it is still dependent of the quality of the segmentation and of the people detection algorithm. In an online situation, for example video surveillance in an airport, storing numerous signatures might also be a problem. However, since the descriptor is based on appearance, the signature database life span would not exceed one day.

In future work, we will focus on building a more complex tracklet signature, using the different positions and postures of the person. Online real-time evaluation of the global tracker will also be used to automatically tune the parameter q depending on the situation. As written in the last section, a trajectory optimization algorithm will also be implemented to improve the tracking metrics results.

Acknowledgments

This work has been supported by the Conseil Regional of Provence-Alpes-Cote d'Azur (PACA).

References

[1] http://www-sop.inria.fr/teams/pulsar/evaluationtool.

[2] S. Bak, E. Corvee, F. Bremond, and M. Thonnat. Multiple-shot Human Re-Identification by Mean Riemannian Covariance Grid. In *AVSS*, Klagenfurt, Austria, Aug. 2011.

[3] C. Beleznai, D. Schreiber, and M. Rauter. Pedestrian detection using GPU-accelerated multiple cue computation. In *CVPR Workshops*, pages 58–65, June 2011.

[4] J. Berclaz, F. Fleuret, and P. Fua. Multiple object tracking using flow linear programming. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, pages 1–8, Dec. 2009.

[5] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance : the CLEAR MOT metrics. In *EURASIP Journal on Image and Video Processing*, volume 2008, pages 1:1–1:10, New York, USA, Jan. 2008.

[6] D. P. Chau, F. Brémond, E. Corve, and M. Thonnat. Repairing People Trajectories based on Point Clustering. In *VISAPP*, pages 449–455, 2009.

[7] D. P. Chau, F. Bremond, and M. Thonnat. A multi-feature tracking algorithm enabling adaptation to context variations. In *ICDP*, London, GB, Nov. 2011.

[8] D. P. Chau, F. Brémond, M. Thonnat, and E. Corvée. Robust Mobile Object Tracking Based on Multiple Feature Similarity and Trajectory Filtering. In *VISAPP*, volume abs/1106.2695, 2011.

[9] D. Conte, P. Foggia, G. Percannella, and M. Vento. Performance Evaluation of a People Tracking System on PETS2009 Database. In *AVSS*, pages 119–126, 29 2010-sept. 1 2010.

[10] E. Corvee and F. Bremond. Haar like and LBP based features for face, head and people detection in video sequences. In *ICVS 2011*, page 10, Sept. 2011.

[11] A. Ellis, A. Shahrokni, and J. Ferryman. PETS2009 and Winter-PETS 2009 results: A combined evaluation. In *Performance Evaluation of Tracking and Surveillance (PETS-Winter)*, pages 1 –8, Dec. 2009.

[12] C.-H. Kuo, C. Huang, and R. Nevatia. Multi-target tracking by on-line learned discriminative appearance models. In *CVPR*, pages 685–692, June 2010.

[13] A.-T. Nghiem, F. Bremond, and M. Thonnat. Controlling Background Subtraction Algorithms for Robust Object Detection. In *ICDP*, London, GB, Dec 2009.

[14] B. Wu and R. Nevatia. Detection and Tracking of Multiple, Partially Occluded Humans by Bayesian Combination of Edgelet based Part Detectors. *Int. J. Comput. Vision*, 75(2):247–266, Nov. 2007.

[15] J. Yang, Z. Shi, P. Vela, and J. Teizer. Probabilistic multiple people tracking through complex situations. In *Performance Evaluation of Tracking and Surveillance workshop at CVPR 2009 (2009)*, number 79-86, pages 1 –8, Dec. 2009.