MULTIPLE OBJECT TRACKING BASED ON SPARSE GENERATIVE APPEARANCE MODELING

Dorra Riahi, Guillaume-Alexandre Bilodeau

LITIV Lab., Polytechnique Montréal

ABSTRACT

This paper addresses multiple object tracking which still remains a challenging problem because of factors like frequent occlusions, unknown number of targets and similarity in objects' appearance. We propose a novel approach for multiple object tracking using a multiple feature framework. The main focus of the proposed method is to build a robust appearance model. The appearance model of an object is built using a color model, a sparse appearance model, a motion model and spatial information. We validated the proposed algorithm on four publicly available videos with comparisons with state-of-the-art approaches. We demonstrate that our algorithm achieves competitive results.

Index Terms— Multiple object tracking, Data association, Tracking by detection, Sparse appearance model

1. INTRODUCTION

Multiple object tracking (MOT) is a fundamental problem to solve for many computer vision applications, such as video surveillance and activity recognition. As a result, research focusing on MOT systems has steadily increased during the past few years. The improvement of MOT systems depends mainly on three factors: detection responses, appearance model of the target and data association between targets and detection responses. Appearance modeling is a crucial process for associating tracks and detections because the observation model can be highly varying and the complex interactions between similar objects may cause ambiguities.

To overcome these difficulties, we introduce a novel MOT method that capitalizes on the strength of multiple cues to build the appearance model. At the core of our approach, we propose a representation that allows distinguishing each target (a target is a region of interest to be tracked). In addition, in agreement with the majority of recent MOT methods, ours is also a tracking by detection method. The data association process is a very challenging task by itself. For simplicity, we use the Hungarian algorithm to associate detection responses frame-per-frame and focus on the appearance model. The contributions of this paper are:

This work was supported by NSERC discovery grant No. 311869-2010

- An assignment method between tracks and candidates (a candidate is a detection response) scored by an affinity function that incorporates four main features which are: color, motion, sparse appearance model and spatial feature.
- An interpolation task for the target position that is based on spatial information. Thus, a target can be tracked even if it is not detected for some time.
- Experiments showing that the proposed approach is applicable to a variety of tracking scenarios and that our approach outperforms recent MOT methods.

As discussed previously, a MOT system can be improved either by improving the detection responses, the data association, or the appearance model. To alleviate the problem of poor detection responses, some recent works add to MOT, model-free single object tracks. They use outputs from an object detector and an object tracker [1] [2] [3] [4]. In [2], the authors use particle filtering responses along with person detector outputs to handle occlusions and missing detections. In a similar spirit, authors in [3], exploit a MOT system that is based on combining outputs from a person detector and a multiple object tracker together. On the other hand, tracking by tracklet was recently proposed [5] [6] [7] [8]. These methods aim to compose tracks from track fragments. For example, in [8], a set of tracklets is built from detection responses. These tracklets are represented using an online learned conditional random field model. In [7], authors proposed a MOT system by linking tracklets into long trajectories by finding a joint optimal assignment between global information (linking tracklets) and local information (linking detection responses). Trajectories are updated iteratively until convergence. Other MOT systems aim to improve the data association process [4] [8] [9]. In [8], authors incorporate a mixed discrete-continuous conditional random field at the data association level.

The approaches described above improve tracking performance, but can be quite complex because of using an object tracker and a graph structure. We believe that creating a robust appearance model should be first addressed. In fact, in the experiment section, we show that a robust appearance model combined with a simple data association strategy can

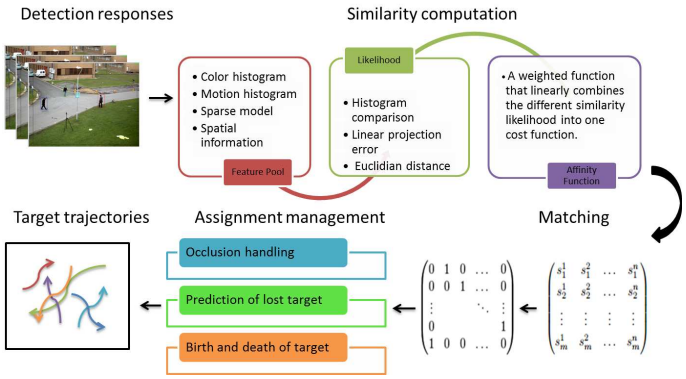

Fig. 1: System Overview.

give good results.

The appearance model is generally based on shape, color appearance [10], motion properties [8] [11] but can be also based on multiple features. For example, in [5], authors use three independent features to model each target which are color histograms, covariance matrices and histograms of gradients (HOG). We take inspiration from these approaches by incorporating a robust appearance model into a simple assignment structure for distinguishing each target over time.

2. PROPOSED METHOD

2.1. Overview of our approach

Our MOT framework is outlined in figure 1. Initially, a set of tracks is created from detection responses in the first frame. Detection responses are independently found in each frame with a trained object detector (detection responses are available with the dataset used). For each frame, an affinity function is calculated based on the similarity between target models and detection responses based on their appearance. A color histogram, a motion histogram, a sparse model and spatial information constitute our features to describe appearance. After calculating the affinity between the set of tracked targets and detection responses, a matching step is achieved by applying the Hungarian algorithm. Finally, we manage the assignments by splitting targets into a set of states based on their assignment and the affinity function.

2.2. Target representation scheme

To obtain a robust appearance model, we use a pool of independent features to represent the targets over time. This choice is justified by the fact that multiple features provide complementary information. We build a global appearance model $M = [H_c, H_m, p, d]$ where H_c is the concatenation of color histograms at each pixel, H_m is the oriented optical flow histogram, p is the probability error of the sparse projection and d is the Euclidean distance between target and candidate center points. Each feature is described below.

Color model (H_c) We use locality sensitive histogram (LSH) to encode the intensity information of a target [12]. The LSH is a recent approach that computes an histogram at each pixel location. The integral histogram at pixel p in the image I is computed as

$$H_p^I(b) = Q(I_p, b) + H_{p-1}^I(b), b = 1, ..., B, \quad (1)$$

where $Q(I_p, b)$ is equal to zero except when intensity value I_p belongs to bin b. The LSH at p is calculated based on the previous histogram computed at pixel $p-1$. To compare LSH histograms, we calculate the Euclidean distance between them.

Motion model(H_m) We use the optical flow to exploit the motion proprieties of the target. The optical flow is calculated according to the original version [13]. To obtain the motion descriptor, we calculate the histogram of oriented optical flow (HOOF). As with the color histograms, we compare the HOOF histograms by calculating the Euclidean distance between them.

Sparse Model (p) Sparse appearance models have attracted a lot of attention in recent years. We adopted and modified the sparse representation technique developed in [14] to fit into our MOT framework. We sparsely projected the detection responses in a template space. A vector of approximate errors of the sparse representation projections is then obtained. It reflects the similarity likelihood between the target sparse model and the candidate (detection response) model. Let $T = \{t_1, t_2, ..., t_n\} \in \mathbb{R}^{d \times n}$ be the set of n target templates (including trivial and non-trivial templates) and $Y = \{y_1, y_2, ..., y_m\}$ be the set of candidates (the detection responses in the current frame). A candidate y_i is sparsely projected into the T template space, that is, y_i can be written as

$$y_i = aT = a_1 t_1 + a_2 t_2 + ... + a_n t_n, \quad (2)$$

where $a = (a_1, a_2, ..., a_n)' \in \mathbb{R}^n$ is the coefficient vector. The observation likelihood for the detection response y_i is

$$p(y_i|x_t) = \frac{1}{\tau} \exp[-\alpha \|y_i - aT\|_2^2], \quad (3)$$

where $||.||_2$ denote the l_2 norm used to solve the minimization problem, x_t is the target model at time t, a is the solution of equation 2, α is a constant, and τ is a normalization factor. The detection response with highest probability (minimum projection error) will be considered as the most similar to the target. An updating step for the template space is then necessary: the weight of each template is increased according to the probability of each candidate.

Spatial model (d) the spatial information is obtained by exploiting the geometric coordinates for both target and the set of candidates. The Euclidean distance is calculated between the target center point and the center points for each candidate:

$$D(i, j) = \sqrt{(i_x - j_x)^2 + (i_y - j_y)^2}, \quad (4)$$

where (i_x, i_y) and (j_x, j_y) are the center coordinate of target and a candidate respectively. Note that this information is taken into account only in the case where there is no occlusion (the target is visible).

2.3. Data association

The matching process plays an important role in our proposed framework. It aims to assign individual observations with tracks by their joint probabilities. Given an affinity function in terms of visual (color and motion model) and contextual properties (spatial and sparse visual model), MOT can be formulated as matching objects across frames. Let $X = \{x_1, x_2, ..., x_n\}$ denote the set of targets currently tracked, $Y = \{y_1, y_2, ..., y_m\}$ be the set of detection responses in frame t and $S = \{s_1, s_2, s_3, s_4\}$ be a set of probability measures at time t for each feature. The affinity function is a quantitative measure that groups all independent feature similarities into one similarity term:

$$f(x_j, y_i) = \sum_k \alpha_k s_k(x_j, y_i), \tag{5}$$

where α_k denotes a weight for each feature and s_k represents the similarity likelihood using the feature number k between the target state x_j and the detection response y_i. After calculating the joint affinity score between the set of targets and candidates, the similarity score is save into $n \times m$ matrix where $f(x_j, y_i)$ is the value in row number j and column number i. Based on the affinity measure, each target will be assigned to one and only one detection response reciprocally. To simplify, the assignment is obtained by applying the Hungarian algorithm. To be more robust to the occlusion problem, we do the assignment in two steps: first we will assign all non-occluded targets with all detection responses, and second we do the assignment between occluded targets and non-assigned detection responses.

2.4. Track management

A MOT system is fundamentally more challenging than a single tracking system due to the variable number of targets over time, heavy interaction between similar objects (in our case the objects are people), unreliable and missing detection responses. We derive a novel strategy to handle occlusions and false detections. If a target is assigned to a detection response, it means that this target is not occluded and it is not in the in/out region. The challenging task is when a target is not assigned or a detection response is not labeled. To handle such difficult cases, each target is defined by a state (active, entering, exiting, occluded).

Birth and death of targets We select manually an in/out area that defines the border of our tracking region. It means that if a person is detected inside this area, it can be set as new track or an exiting target. At a frame t, if a target is not assigned (or assigned with a very low affinity score) and the target is detected in the in/out area during k frames, then this track state will be set as exiting target and then it will be deleted from the set of tracks. On the other hand, if a detection response is not assigned and it is detected in the in/out area, we can hypothesize that a new person is entering in the field of view of the camera. In this case, the track state will be set as entering and we should add this new target to the set of tracks.

Occlusion handling In order to handle occlusions, we added a state that indicates if the target is visible or occluded. A non-assigned target means that this target is occluded by other objects, or that the detection of the target is missing. If the target is not assigned or its affinity function is smaller than a defined threshold, the assignment will be deleted and the track state will be occluded target. During the occlusion time, we should stop tracking the person. Once the occluded target is reassigned again with a significant affinity score, the target's state will be set as active target.

Interpolation of track of lost targets We use an interpolation step in order to handle false positive and false negative detection responses. The interpolations are achieved based on the motion vector that describes the history of movement of the target over time. We can interpolate lost target only if it reappears. In pedestrian tracking, we can assume that the person walks in a linear motion. To find the last unknown position of the target, we linearly interpolate the coordinates of the target between the last position (just before the loss of the target) and the current position.

3. EXPERIMENTAL RESULTS

We validated our proposed method on a variety of challenging sequences: TUD Campus [15], TUD Crossing [15], TUD Stadtmitte [16] and PETS09 S2-L1 [16]. They have the following properties. They show walking pedestrians outside, so lighting conditions can be very poor. Furthermore, because of the camera angle, people get very small when they are far from the camera making their tracking more challenging (PETS09 video). In TUD dataset, targets are frequently occluding each other (heavy inter-object occlusion) and are occluded by static objects. We use the detection responses provided with the videos. For each detected person, a classification confidence value is provided.

Tracking performance is evaluated using the widely used CLEAR MOT metrics [17]. They include an accuracy score called (MOTA) that combines false positive, missed targets and identity switch errors and a precision score called (MOTP) that is the average distance between ground truths and predicted targets. In addition, the CLEAR MOT metrics include: false negative (FN), false positive (FP) and the numbers of identity switches (IDS).

Dataset	Method	MOTA	MOTP	FN	FP	IDS
TUD-CAMPUS	Proposed	***78.18%***	***69%***	0%	13%	0
	[20]	72%	***74%***	25 %	2%	1
	[1]	***73.3%***	67%	26.4%	0.1%	2
TUD-CROSSING	Proposed	***78.30%***	66%	1.4%	8.38%	7
	[20]	72%	***76%***	26%	1%	7
	[1]	***84.30%***	***71%***	14.10%	1.4%	2
TUD-STADTMITTE	Proposed	***67%***	57.26%	26%	5.74%	22
	[2]	60.5%	***65.8%***	-	-	7
	[16]	56.2%	61.6%	-	-	15
	[9]	***63%***	***73%***	-	-	-
PETS09-S2-L1	Proposed	***84%***	***66%***	13%	2%	35
	[18]	75.9%	53.8%	-	-	-
	[1]	79.7%	56.3%	-	-	-
	[2]	***80.3%***	***76%***	-	-	15
	[21]	60%	***66%***	-	-	-
	[19]	70.36%	-	-	-	-

Table 1: Comparaison of results on TUD and PETS09 dataset. Best method in ***red*** and second best in ***blue***

3.1. Quantitative comparison

We compared our method with state-of-the-art methods when available (table 1). On TUD-Campus, we outperform approache [1] which uses an online model-free tracker besides the detector outputs. On PETS09-S2-L1 video sequence, our MOTA is higher than in the previous results. Interestingly, our algorithm outperforms the method of Breitenstein et al. [1] that uses outputs from particle filter tracker and HOG detector. This shows the robustness of our appearance model. Furthermore, we perform better than Yang and al. method [18] which includes background subtraction to handle occlusion. It is possible to observe that our multiple tracking system accuracy is higher than Gustavo and al. approach [19] by around 14% even if they use multiple patches to represent their appearance model.

Although, due to the variation of the targets' appearance and many occlusions, the tracking process is more challenging in PETS09-S2-L1, we succeed in getting the best accuracy compared the other recent approaches. This success shows that our tracking management is robust even if simple. For TUD Stadtmitte, we also outperform Milan and al. [2] MOT system based on continuous optimization and the method developed in [16] based on discrete-continuous conditional random field for the MOTA metric. MOTP is a little lower than these methods because of a higher number of IDswitch. This is the drawback of our MOT approach. This is explained by the fact that if we are not sure that the assignment is correct, we add a new track.

The results presented in table 1 emphasis the fact that the use of a robust appearance model with a simple technique of data association can achieve better results. The robustness of our appearance model is coming from the use of a sparse representation model in addition to other independent features.

Figure 2 depicts an example of the results of our approach on PETS09-S2-L1 dataset which is the most difficult video. We can see that our algorithm can easily handle heavy occlusion between people in a case of moderately crowded scenes

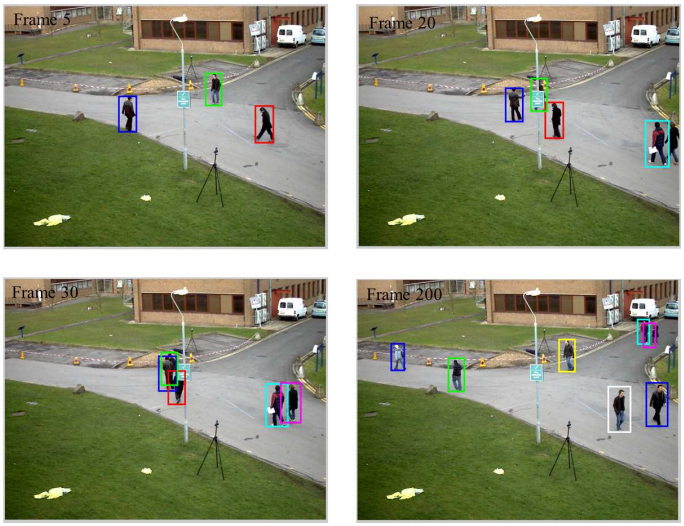

Fig. 2: Results for the PETS2009 dataset. Top left: frame 5,top right: frame 20, bottom left: frame 30 and bottom right: frame 200

(see blue and green targets). Even if the appearance changes, our MOT approach can identify the targets (see blue target). Figure 2 also illustrates how our proposed algorithm is robust to scale change (see cyan and magenta targets).

4. CONCLUSION

The proposed MOT method can be divided into two components: the appearance model and the tracking process. For the first task, we use multi-cue fusion to get more information on target appearance to improve the robustness of our framework. To handle important MOT problem, we exploit a strategy that can lessen the problem of missing detection responses by interpolation of the lost track. Our main contribution is thus to propose a simple and robust framework that combines more features for multiple object tracking. The proposed method is compared to several state-of-the-art approaches, which demonstrate the benefits of our method. Our method is competitive on all tested videos.

5. REFERENCES

[1] Michael D Breitenstein, Fabian Reichlin, Bastian Leibe, Esther Koller-Meier, and Luc Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *Pattern Analysis and Machine Intelligence(PAMI), IEEE Transactions on*, vol. 33, no. 9, pp. 1820–1833, 2011.

[2] Anton Andriyenko and Konrad Schindler, "Multi-target tracking by continuous energy minimization," in *CVPR*. IEEE, 2011, pp. 1265–1272.

[3] Xu Yan, Xuqing Wu, Ioannis A Kakadiaris, and Shishir K Shah, "To track or to detect? an ensemble framework for optimal selection," in *ECCV*, pp. 594–607. Springer, 2012.

[4] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *CVPR*. IEEE, 2008, pp. 1–8.

[5] Cheng-Hao Kuo, Chang Huang, and Ram Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," in *CVPR*. IEEE, 2010, pp. 685–692.

[6] Bing Wang, Gang Wang, Kap Luk Chan, and Li Wang, "Tracklet association with online target-specific metric learning," in *CVPR*. IEEE, 2014, pp. 1234–1241.

[7] Shun Zhang, Jinjun Wang, Zelun Wang, Yihong Gong, and Yuehu Liu, "Multi-target tracking by learning local-to-global trajectory models," *Pattern Recognition*, 2014.

[8] Bo Yang and Ramakant Nevatia, "Multi-target tracking by online learning a crf model of appearance and motion patterns," *International Journal of Computer Vision(IJCV)*, vol. 107, no. 2, pp. 203–217, 2014.

[9] Aleksandr V Segal and Ian Reid, "Latent data association: Bayesian model selection for multi-target tracking," in *ICCV*. IEEE, 2013, pp. 2904–2911.

[10] Siyu Tang, Mykhaylo Andriluka, and Bernt Schiele, "Detection and tracking of occluded people," *International Journal of Computer Vision(IJCV)*, pp. 1–12, 2012.

[11] Alper Yilmaz, Omar Javed, and Mubarak Shah, "Object tracking: A survey," *ACM computing surveys (CSUR)*, vol. 38, no. 4, pp. 13, 2006.

[12] Shengfeng He, Qingxiong Yang, Rynson WH Lau, Jiang Wang, and Ming-Hsuan Yang, "Visual tracking via locality sensitive histograms," in *CVPR*. IEEE, 2013, pp. 2427–2434.

[13] Berthold K Horn and Brian G Schunck, "Determining optical flow," in *1981 Technical Symposium East*. International Society for Optics and Photonics, 1981, pp. 319–331.

[14] Chenglong Bao, Yi Wu, Haibin Ling, and Hui Ji, "Real time robust l1 tracker using accelerated proximal gradient approach," in *CVPR*. IEEE, 2012, pp. 1830–1837.

[15] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," in *CVPR*, June 2008, pp. 1–8.

[16] Anton Milan, Konrad Schindler, and Stefan Roth, "Detection-and trajectory-level exclusion in multiple object tracking," in *CVPR*. IEEE, 2013, pp. 3682–3689.

[17] Bernardin Keni and Stiefelhagen Rainer, "Evaluating multiple object tracking performance: the clear mot metrics," *EURASIP Journal on Image and Video Processing*, vol. 2008, 2008.

[18] Jun Yang, Patricio A Vela, Zhongke Shi, and Jochen Teizer, "Probabilistic multiple people tracking through complex situations," in *11th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, 2009.

[19] Gustavo Führ and Cláudio Rosito Jung, "Combining patch matching and detection for robust pedestrian tracking in monocular calibrated cameras," *Pattern Recognition Letters*, vol. 39, pp. 11–20, 2014.

[20] Dorra Riahi and Guillaume-Alexandre Bilodeau, "Multiple feature fusion in the dempster-shafer framework for multi-object tracking," in *Computer and Robot Vision (CRV)*. IEEE, 2014, pp. 313–320.

[21] Jerome Berclaz, Francois Fleuret, and Pascal Fua, "Robust people tracking with global trajectory optimization," in *CVPR*. IEEE, 2006, vol. 1, pp. 744–750.