Handling Inter-object Occlusion for Multi-object Tracking Based on Attraction Force Constraint

Yuke Li(✉), Isabelle Bloch, and Weiming Shen

State Key Laboratory of LIESMARS, Wuhan University, Wuhan, China
{leesunfreshing,wmshen66}@gmail.com, isabelle.bloch@telecom-paristech.fr

Abstract. This paper presents a novel social interaction relation, attraction (interaction that would lead to occlusion for inter-object) for multi-object tracking to handle occlusion issue. We propose to build attraction by utilizing spatial-temporal information from 2D image plane, such as decomposed distance between objects. Then pairwise attraction force is obtained by the modeled attraction. Lastly, the attraction force is used to improve tracking when hierarchical data association performs. To meet requirements of practical application, we have our method evaluated on widely used PETS 2009 datasets. Experimental results show that our method achieves results on par with, or better than state-of-the-art methods.

Keywords: Attraction force · Occlusion handling · Multi-object tracking

1 Introduction

Inter-object occlusion is one of the most difficult task to deal with in object tracking field. This issue could be explained by the spatial-temporal information for the objects, that are involving occlusion is quite different from those are not. However, most of these approaches ignore that spatial-temporal information is not exploited sufficiently.

Many research have been accomplished great achievement w.r.t. occlusion handling. In [1,7,17], the authors focus on focus on the appearance change while occlusion happens. [12] propose to utilize scene knowledge to solve objects missing caused by occlusion. Nevertheless, most of them neglect the spatial-temporal information when occlusion happens. By contrast, the social force interaction among multi-object [6], which is based on exploring spatial-temporal information, provides a different perspective for multi-object tracking. Whereas none of the research in such a field considers that, inter-object occlusion is caused by social force interaction. For example, [11] and [13] use the interaction to predict objects location, without considering occlusion between objects.

Intuitively, the spatial-temporal information for objects that are involving occlusion is different from those are not. For instance, the distance for those

M. Kamel and A. Campilho (Eds.): ICIAR 2015, LNCS 9164, pp. 520–527, 2015.
DOI: 10.1007/978-3-319-20801-5_58

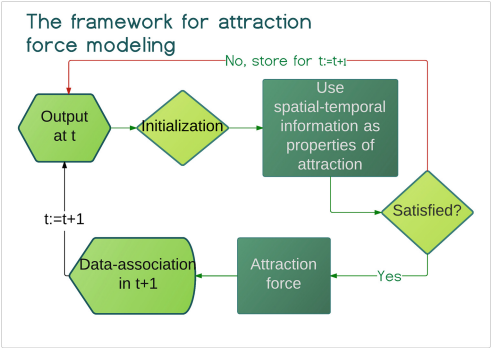

Fig. 1. Our framework of attraction force modeling. After initialization, the spatial-temporal information, such as distance, is utilized to decide whether there is attraction. The final step would be to integrate the attraction force into the tracking scheme.

occluded objects would be relative closer. Similar rationale has been employed lately for scene understanding [4]. The idea of considering the global scene spatial-temporal information has been receiving great attention in the field of more complex activity recognition [3] as well. Motivated by this intuition and based on the observation above, in this work we present a framework of inter-object occlusion handling for multi-object tracking based on attraction force (social force that may lead to occlusion between objects). The framework of our attraction force modeling is shown in Fig. 1.

The contributions of our work[1] are summarized as follows: 1. We extend the concept of social force by building attraction force. Attraction force is particular for the situations that would lead to inter-object occlusion. This model is completely based on 2D image plane information without any scene knowledge, such as camera calibration, etc. By utilizing change of distance between objects, the relative velocity as well, we propose that attraction force suggests information of occlusion between objects in next frame. 2. A novel occlusion handling method is proposed. Our approach focus on dealing with occlusion in data-association level. Attraction force is used as penalty to optimize final association score. The authors in [2] utilizes a similar rationale, But our method differs in occlusion modeling and data association framework

The reminder of this paper is organized as follows: Firstly, hierarchical tracking-by-detection framework is discussed in Sect. 2. Section 3 focus on modeling attraction force utilizing spatial-temporal information, and handling occlusion based on attraction force in hierarchical data-associations, followed by a set of detailed experimental results and analysis in Sect. 4. Finally, we conclude in Sect. 5.

[1] This work is performed when the first author was with Institut Mines Télécom, Paris. The author would like to thank Prof. Isabelle Bloch, Dr. Ling Wang and Dr. Henrique Morimits for meaningful discussion and very helpful suggestions.

2 Hierarchical Tracking-by-Detection

Online tracking-by-detection approach combines discriminative [14] and generative methods [10] for multi-object tracking. Such a method treats frame by frame data association as pair-wise assignment problem, that matches the detection with tracking results. In our work, hierarchical data-association method is adopted. Assuming in t frame, all the detection inputs are taken as one of detection division $\mathcal{DE}$, and tracking results are taken into target division $\mathcal{TR}$. Candidates $\mathcal{CA}$ is the subset of $\mathcal{DE}$, which is used to represent new objects appear in the scene. To sum up, we have $\mathcal{DE} + \mathcal{TR}$ as input for every frame. Regarding birth and death of tracker, we follow the same procedure in [16], which is the new tracker is generated from $\mathcal{CA}$. The data-association would be performed between $\mathcal{DE}$ and $\mathcal{TR}$. Noted that since the $\mathcal{CA}$ may contain false positive, we follow the procedure by [2], tracker will be generated when one candidate is matched for at least 2 consecutive frames.

To assign correct detection to correct tracking result, one matching score by computing likelihood between detections and tracking results is used. The matching score (M) includes several components, in our case, we use

$$M = Pos \cdot Size \cdot App \tag{1}$$

where $Pos = \mathcal{N}(0, dis(P_{de} - P_t r))$, with (P_*) is position of detector and tracker in current frame, $Size = \mathcal{N}(0, \frac{size_{de} - size_{tr}}{size_{tr}})$ with $(size_*)$ is the size of detector and tracker. $\mathcal{N}$ is Gaussian distribution with zero mean. For the appearance, we employ Hellinger distance the HSV color histogram. It consists on computing the histogram of both detector and tracker on the HSV color space. In order to deal with situation such as illumination changing and occlusion, we keep the color histogram information of first frame and last frame that object has correctly tracked . After having those matching score, Hungarian algorithm [9] provides the best match.

For each object, the tracking result and its matched detection is output, if there is one; otherwise, only the tracking result is used as output instead.

3 Inter-object Occlusion Handling Strategy Based on Attraction Force

Even hierarchical tracking-by-detection tackles occlusion implicitly by using color histogram from first and last frame. The matching score cannot always give the best result when inter-object occlusion occurs, since detector may not always find the right object under occlusion. In this section, we will present our method to handle occlusion by detailedly analysis and model attraction force.

3.1 Initialization of Attraction Force

We manually set search region as square shape to eliminate the objects that are too far away to have attraction. The size of one side of the search region

is considered as twice as the height of the object. Euclidean distance between center point of bounding boxes of objects is employed to estimate the distance among objects. Only the objects are within the search region of other objects, and without any occlusion are initialized for attraction force. Additionally, If the overlapping area of two bounding boxes is more than 40 %, we consider the spatial information is invalid in order to avoid potential errors. Furthermore, the example that to deal with size of the object may lead to information lost. For instance, x^1 and x^2 is object with bigger and smaller size respectively. When we consider attraction of x^2, besides of all the objects within the search region, we need enumerate attraction of x^1. If we find out attraction between x^1 and x^2 for x^1, this information is stored and taken into account for x^2.

3.2 Attraction Analysis

Intuitively, inter-object occlusion would only happen for objects moving towards each other from at least one axis from image plane. Let us start with the situation such that two objects annotated as x^i and x^j walk towards each other , and there is an attraction between them. The following equation is used to describe this situation:

$$\begin{cases} {D^X}_{t-1}(x^i, x^j) - {D^X}_t(x^i, x^j) > 0 \\ {D^Y}_{t-1}(x^i, x^j) - {D^Y}_t(x^i, x^j) > 0 \end{cases} \tag{2}$$

where ${D^X}_{t-1}(x^i, x^j)$ and ${D^Y}_{t-1}(x^i, x^j)$ is the Euclidean distance between x^i and x^j in X and Y axis, at time $t-1$ and t. In order to avoid some confused ambiguities lead by general distance, decomposed distance in X and Y axis is employed. For instance, when two pedestrian are passing by each other from image plane, no occlusion will be observed. Nevertheless utilizing the general distance is hard to distinguish whether these two objects are passing by or having occlusion. This reason remains in the following.

Equation 2 implies two things: firstly, x^i and x^j are closer; besides, the relative displacement of x^i and x^j between two subsequent frame could present as relative velocity between these two objects, and Eq. 2 indicates that they tend to meet each other.

Two objects are moving towards from only one axis, X axis is used as paradigm. Two cased is considered for this situation. The first one is no distance change in Y axis, and the second one is repelling from Y axis. To asses if there is attraction, we still rely on the distance information, with aiding by the size of the object. First case could be described by

$$\begin{cases} {D^X}_{t-1}(x^i, x^j) - {D^X}_t(x^i, x^j) > 0 \\ {D^Y}_{t-1}(x^i, x^j) - {D^Y}_t(x^i, x^j) = 0 \\ {D^Y}_t(x^i, x^j) < 0.5(H_{x^i} + H_{x^j}) \end{cases} \tag{3}$$

and second case is

$$\begin{cases} {D^X}_{t-1}(x^i, x^j) - {D^X}_t(x^i, x^j) > 0 \\ {D^Y}_{t-1}(x^i, x^j) - {D^Y}_t(x^i, x^j) < 0 \\ \frac{0.5(H_{x^i} + H_{x^j}) - {D^Y}_{t,x^i}(x^j)}{{V^Y}_{t,x^i}(x^j)} > \frac{-0.5(W_{x^i} + W_{x^j}) + {D^X}_{t,x^i}(x^j)}{{V^X}_{t,x^i}(x^j)} \end{cases} \tag{4}$$

where H_{x^*} and W_{x^*} is height and width of the x^i and x^j respectively. and ${V^*}_{t,x^i}(x^j)$ is the relative velocity between x^i and x^j..

Two objects have attraction or not judging by the size information.

Similar situation for Y axis is symmetrical to the situation described by Eq. 3 and Eq. 4 (simply switch X and Y, height and width).

3.3 Attraction Force

Attraction modeled previously is utilized as properties for attraction force. Noted only objects satisfy one of these properties, will be taken into account for attraction force.

The attraction force is modeled as:

$$\begin{cases} F_{att}^{X,t}(x^i, x^j) = I \cdot \left(1 - exp^{-\left|{V^X}_{t,x^i}(x^j)\right| \cdot (\alpha - {D^X}_t(x^i, x^j))}\right) \\ F_{att}^{Y,t}(x^i, x^j) = I \cdot \left(1 - exp^{-\left|{V^Y}_{t,x^i}(x^j)\right| \cdot (\alpha - {D^Y}_t(x^i, x^j))}\right) \end{cases} \tag{5}$$

I is indicator function that equals one if there is attraction between x^i and x^j based on the explanation of previous sections, equals zeros if otherwise. α equals to the height of x^i. $\alpha - {D^X}_t(x^i, x^j) > 0$ makes sure x^j is within the search region of x^i. $\left|{V^X}_{t,x^i}(x^j)\right|$ is the absolute of relative velocity of x^i and x^j, which is defined by object state (Subsect. 4.3).

3.4 Occlusion Handling Strategy

In this work, all inter-object occlusion relation (only the matching score $> \tau$ is taken as occlusion, where τ is the threshold manually set) between objects will be enumerated within the same division, only the matching score of at least two detections and two targets are considered connected, are treated as occlusion group. As we have already pointed in previous subsection, attraction force based occlusion rationale is to predict occlusion for $t+1$. Therefore the data-association in this section performs in $t+1$ as well. Assuming that d_m, tr_n are within the same occlusion group, we propose

$$\hat{M} = \arg\max_{m,n} \sum_{m,n} M(d_m, tr_n) - \hat{F_{att}^t} \tag{6}$$

for optimizing matching score. Where $\hat{F}$ is overall attraction force in this occlusion group, and used to penalize the occlusion. $\hat{F_{att}^t}$ is only considered once for each pair. For example, if attraction force $F_{att}^t(x^h, x^k)$ exists between occlusion group which is comprised of object x^h and x^j, $\hat{F_{att}^t} = F_{att}^t(x^h, x^k)$.

4 Experiments

4.1 Datasets

To better evaluate the capability of our method, We have our approach tested on widely used PETS2009 benchmark provided by [15].

The most challenges of this dataset is frequently occlusion caused by dynamic pedestrians movement. We also run experiments on more challenged S2L2 datasets, which more pedestrian presents in the scene.

To achieve fair comparison score, we use the goundtruth provided by [15], where all the person occurring in the scene have been annotated.

4.2 Metrics

To measure performance, the CLEAR MOT metrics [8] is adopted. The metrics include: 1. Multiple Object Tracking Accuracy (MOTA, higher value is better) returns a accuracy score; 2. Multiple Object Tracking Precision (MOTP, lower value is better), which consider intersection union of bounding boxes; 3. Mostly Tracker (MT) and 4. Mostly lost (ML). MT and ML is not used for our evaluation for PETS2009 S2L2 dataset, because most of the methods in our comparison do not provide such a value. The procedure provided by [9] is adopted, in which the results will be re-evaluated by 2D matching protocol.

4.3 Experiment Settings

Performing tracking, Kalman filter is employed. The entire object state is defined as $\{\mathcal{X}, \mathcal{Y}, \mathcal{H}, \mathcal{W}, {V^X}_t, {V^Y}_t\}$, where $\mathcal{X}, \mathcal{Y}$ is the position in X axis and Y axis respectively, $\mathcal{H}, \mathcal{W}$ is the height and width of x^i, ${V^X}_t, {V^Y}_t$ is the velocity w.r.t each axis. The noise of them are manually set as $\mathcal{N}(0, 10)$ and $\mathcal{N}(0, 5)$ respec tively, Δt is the time between 2 consecutive frames.

The noises of them are both manually set as $\mathcal{N}(0, 10)$ and $\mathcal{N}(0, 5)$ respectively. $V^X = V^Y = 0$ at the frame that tracker is initialized.

Similar to [12], DPM (Deformable Part based Model detector)[5] is utilized to generate detection input. In additional, following the settings presented in [16], the false positives of detection is removed by the size.

4.4 Results and Analysis

Figure 2 depicts our results for exemplar two consecutive frames. Table 1 illustrates the quantitative comparison of our method and state-of-the-art online tracking approaches.

Table 1 compares the performance of the tracker on PETS2009 S2L1 dataset. Utilizing scene knowledge, camera calibration for instance, makes the approach of [12] outperforms to other methods, however, our approach performs favorably compared to most of online trackers. Our novel occlusion handling improves the capability to deal with occlusion of the tracker. For the tracking method employ similar hierarchical data-association scheme [16], our occlusion handling makes the tracker more robust. Thus, we achieve better MOTA score and significant improved MOTP score.

Considering more challenged PETS2009 S2L2 sequences, our method provides better scores in both MOTA and MOTP than most of other approaches.

Fig. 2. The experimental results of our method for exemplar two consecutive frames, on PETS2009-S2L1 (the first and the second on the left) and PETS2009-S2L2 (the first and the second on the right) respectively. Our method shows good capability to handle occlusion. The details of experiments presents in Sect. 4.

Both datasets confirm that the proposed method are beneficial by employing occlusion handling method. Comparing with other methods only consider appearance under occlusion [2,12,16], spatial-temporal information could be more reliable. Furthermore, the experimental results suggest that, employing scene knowledge [12] may lead to further improvement of tracking performance.

Table 1. Comparison of different online tracking methods.

PETS2009-S2L1	MOTA [%]	MOTP [%]	MT [%]	ML [%]	PETS2009-S2L2	MOTA [%]	MOTP [%]
Proposed method	93.6	71.3	100.0	0.0	Proposed method	68.2	60.7
Breitenstein et al. [2]	79.7	56.3	-	-	Breitenstein et al. [2]	50.0	56.3
Possegger et al. [12]	98.1	80.5	100.0	0.0	Possegger et al. [12]	66.0	64.8
Jianming et al. [16]	93.4	68.2	100.0	0.0	Jianming et al. [16]	66.7	58.6

5 Conclusion

A novel occlusion handling method based on attraction force is proposed.By detailed analysis every possible situation would lead to attraction, occlusion handling is performed in data-association level. The experimental results show that our method could be comparable with, even better than state-of-the-art.

References

1. Andriyenko, A., Roth, S., Schindler, K.: An analytical formulation of global occlusion reasoning for multi-target tracking. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 1839–1846. IEEE (2011)

2. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Online multiperson tracking-by-detection from a single, uncalibrated camera. IEEE Trans. Pattern Anal. Mach. Intel. **33**(9), 1820–1833 (2011)
3. Chang, X., Zheng, W.S., Zhang, J.: Learning Person-Person interaction in collective activity recognition. IEEE Trans. Image Proces. **24**(6), 1905–1918 (2015)
4. Choi, W., Chao, Y.W., Pantofaru, C., Savarese, S.: Understanding indoor scenes using 3d geometric phrases. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 33–40. IEEE (2013)
5. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part based models. IEEE Trans. Pattern Anal. Mach. Intel. **32**(9), 1627–1645 (2010)
6. Helbing, D., Molnar, P.: Social force model for pedestrian dynamics. Phys. Rev. E **51**(5), 4282 (1995)
7. Hua, Y., Alahari, K., Schmid, C.: Occlusion and motion reasoning for long-term tracking. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 172–187. Springer, Heidelberg (2014)
8. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol. IEEE Trans. Pattern Anal. Mach. Intel. **31**(2), 319–336 (2009)
9. Kuhn, H.W.: The hungarian method for the assignment problem. In: 50 Years of Integer Programming 1958–2008, pp. 29–47. Springer, Heidelberg (2010)
10. Nummiaro, K., Koller-Meier, E., Van Gool, L.: An adaptive color-based particle filter. Image Vis. Comput. **21**(1), 99–110 (2003)
11. Pellegrini, S., Ess, A., Schindler, K., Van Gool, L.: You'll never walk alone: Modeling social behavior for multi-target tracking. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 261–268. IEEE (2009)
12. Possegger, H., Mauthner, T., Roth, P.M., Bischof, H.: Occlusion geodesics for online multi-object tracking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014)
13. Ramanathan, V., Yao, B., Fei-Fei, L.: Social role discovery in human events. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2475–2482. IEEE (2013)
14. Tang, S., Andriluka, M., Milan, A., Schindler, K., Roth, S., Schiele, B.: Learning people detectors for tracking in crowded scenes. In: 2013 IEEE International Conference on Computer Vision (ICCV), pp. 1049–1056. IEEE (2013)
15. Yang, B., Nevatia, R.: Multi-target tracking by online learning of non-linear motion patterns and robust appearance models. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1918–1925. IEEE (2012)
16. Zhang, J., Presti, L., Sclaroff, S.: Online Multi-person Tracking by Tracker Hierarchy. In: 2012 IEEE Ninth International Conference on Advanced Video and Signal-Based Surveillance (AVSS), pp. 379–385, September 2012
17. Zhang, T., Jia, K., Xu, C., Ma, Y., Ahuja, N.: Partial occlusion handling for visual tracking via robust part matching. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1258–1265. IEEE (2014)