Inter-Person Occlusion Handling with Social Interaction for Online Multi-Pedestrian Tracking*

Yuke LI\(^{a)}\), Member and Weiming SHEN\(^{\dagger)}\), Nonmember

SUMMARY  Inter-person occlusion handling is a critical issue in the field of tracking, and it has been extensively researched. Several state-of-the-art methods have been proposed, such as focusing on the appearance of the targets or utilizing knowledge of the scene. In contrast with the approaches proposed in the literature, we propose to address this issue using a social interaction model, which allows us to explore spatio-temporal information pertaining to the targets involved in the occlusion situation. Our experimental results show promising results compared with those obtained using other methods.

key words: image processing, computer vision, inter-person occlusion, online tracking

1. Introduction

There has been increasing need for applications that serve security-related purposes, such as robotics and surveillance. To this end, the application of multi-object tracking has received more attention because it plays a critical role in this field. One of the most challenging issues is the handling of inter-person occlusion, and it is one of the key aspects considered for tracking. Many reported methods\(^{1]}\)–\(^{3]}\) consider the changes in appearance during occlusion. Others \(^{4]}\) focus on solving the issue of missed targets by utilizing scene knowledge caused by occlusion. In contrast with these state-of-the-art approaches, research on social-force interactions use spatio-temporal information for computer-vision tasks such as tracking\(^{5]}–^{7]}\), activity recognition\(^{8]}\), \(^{9]}\), and abnormality detection\(^{10]}\). However, most of them ignore the fact that inter-person occlusion may be caused by social interaction.

Intuitively, based on observations from an image plane, persons involved in occlusion are expected to present different spatial-temporal patterns from those who are not. For example, the distance between two occluded people as observed from an image plane would be relatively closer than two people who are not being occluded. Based on the observations above, in this paper, we aim to find an occlusion-handling solution for online tracking. Further, we exploit the spatio-temporal information and propose a solution for inter-person occlusion handling.

To differentiate between the contributions of our previous work\(^{11]}\), the following aspects are extended: 1. The attraction force, which is derived from the concept of social interaction\(^{5]}\), \(^{8]}\), \(^{12]}\), is introduced to deal with occlusion. Social interaction is an intuitive concept based on people’s behavior. People would like to choose different paths in order to avoid collision, for instance. The state-of-the-art methods usually consider modeling social interaction as an energy function to express its power. These energy functions would be used as a feature to explore its influence of people’s actions, such as walking together or departing. Hence our model is particularly suitable for situations involving inter-person occlusion. This model is based entirely on two-dimensional (2D) image plane spatio-temporal information without any scene knowledge, such as camera calibration. By utilizing the changes in the distance between pedestrians as well as their relative velocities, our proposed model suggests that there is occlusion information among pedestrians in the next frame. 2. Occlusion is handled in the data-association stage. The attraction force is used as a constraint to determine whether there will be occlusion among targets. To this end, the optimized final-association score is capable of serving this purpose. In \(^{13]}\), a similar rationale has been used. However, our method differs from that study in terms of the occlusion modeling and data-association scheme. Figure 1 shows the pipeline of our pro-

---

*Copyright © 2016 The Institute of Electronics, Information and Communication Engineers*
posed tracking framework.

The remainder of this paper is organized as follows: First, in Sect. 2, we introduce an online-tracking framework that is based on a hierarchical tracking-by-detection framework. In Sect. 3, we discuss how to utilize two specific models derived from a social interaction force, attraction and grouping, based on spatial-temporal information. The occlusion handling strategy is in the data association state, followed by a set of detailed experimental results and analysis in Sect. 4. Finally, in Sect. 5, we summarize this paper.

2. Spatial Temporal Analysis for Inter-Person Occlusion

Observing from the image plane, inter-person occlusion would occur only when persons are walking towards each other, or when they are walking together. In this section, in order to parse inter-person occlusion, we present a detailed analysis of spatial temporal information with respect to attraction and grouping.

2.1 Analysis for Attraction

The most intuitive information that can be observed from the image plane is that the person is moving. It is generally measured by decomposing the X- and Y-axes, respectively, and then combining them together. We also employed this strategy for our analysis.

In this paper, the definition of an attraction is similar to that in [14], which is taken people’s moving velocity, current position et al. to predict if people would eventually meet. It is further modified to serve the occlusion-handling purpose.

We start with the case where two pedestrians annotated as $x^i$ and $x^j$ walk towards each other (see Fig. 2(a)). The following equation is used to describe this situation:

$$\begin{align*}
\Delta D^X_{t-1,i}(x^i, x^j) & = D^X_{t-1,i}(x^i, x^j) - D^X_t(x^i, x^j) > 0 \\
\Delta D^Y_{t-1,i}(x^i, x^j) & = D^Y_{t-1,i}(x^i, x^j) - D^Y_t(x^i, x^j) > 0
\end{align*} \tag{1}$$

where $D^X_{t-1,i}(x^i, x^j)$ and $D^Y_{t-1,i}(x^i, x^j)$ are the Euclidean distances between $x^i$ and $x^j$ in the X- and Y-axis, at time $t-1$ and $t$, respectively. In order to avoid some confused ambiguities, lead by general distance, decomposed distance in X and Y axis is employed. For instance, two pedestrians are approaching each other from the X-axis. No occlusion will be observed if the distance in the Y-axis is greater than $0.5(H_{x^i} + H_{x^j})$ (this will be explained below). Nevertheless, utilizing the general distance, it is hard to distinguish whether these two pedestrians are passing each other or having an occlusion. This is assumed for the remainder of this paper.

Equation (1) implies two things. First, the distance between $x^i$ and $x^j$ decreases for both axes; besides, the relative displacement of $x^i$ and $x^j$ between the two subsequent frames could present as the relative velocity between the two pedestrians. Equation (1) also indicates that they tend to meet each other.

When two pedestrians are moving towards each other from X axis, two cases can be considered. The first is where there is no change in the distance in the Y-axis, and the second one is when it is repelled from the Y-axis.

To assess whether there is an attraction, we still rely on the distance information, aided by the size of the pedestrian. The first case could be described by (see Fig. 2(b))

$$D^Y_t(x^i, x^j) < 0.5(H_{x^i} + H_{x^j}) \text{ while }$$

when two pedestrian are moving towards from one axis, X axis is studied as a paradigm. Two cases are considered for this situation. The first one is no distance change in Y axis, and the second one is repelling from Y axis.

To assess whether there is attraction, we still rely on the distance information, aiding with the size of the pedestrian. The first case could be described by (see Fig. 2(b))

$$D^Y_t(x^i, x^j) < 0.5(H_{x^i} + H_{x^j}) \text{ while }$$

Assume that if $D^Y_{t}(x^i, x^j) \geq 0.5(H_{x^i} + H_{x^j})$, there is no occlusion between $x^i$ and $x^j$.

The second case (Fig. 2(c)) is

$$\frac{V_{x^i,i} \cdot D^X_t(x^i, x^j)}{V_{x^i,i}^2} > -0.5(W_{x^i} + W_{x^j}) + D^X_t(x^i, x^j)$$

where $H_{x^i}$ and $W_{x^i}$ is height and width of the $x^i$ and $x^j$ respectively, and $V_{x^i,i}$ is the relative velocity between $x^i$ and $x^j$. 

In Eq. (3)
serves the same purpose of $D^V_i(x^t, x^u) < 0.5(H_r + H_c)$ in Eq. (2), which serves to determine whether attraction existed or not aiding by size information. Given the relative velocity $V^r_{i,x}(x^t)$, we assume $x^t$ and $x^u$ will meet in X axis before they are totally separated in Y axis.

For the Y-axis, there is a similar situation that is symmetrical to the case described by Eq. (2) and Eq. (3) (simply switch the X- and Y-axes, as well as the height and width).

The above analysis model serves to provide properties of the attraction force model. When pedestrians satisfy one of these properties, it is considered that an attraction exists. According to [5], [14], [15], the rationale behind social interaction is to build an energy function to express the power from it. Thus, the attraction force model is proposed to further quantify the influence as follows:

\[
\begin{align*}
F^V_{\text{att}}(x^t, x^u) &= I \cdot (1 - e^{-|V^r_{i,x}(x^t, x^u)|/(\alpha - D^V_i(x^t, x^u))}) \\
F^V_{\text{att}}(x^t, x^u) &= I \cdot (1 - e^{-|V^r_{i,x}(x^t, x^u)|/(\alpha - D^V_i(x^t, x^u))})
\end{align*}
\]

(4)

I is the indicator function, which equals one if there is an attraction between $x^t$ and $x^u$ based on the explanation given in the previous sections; otherwise, it is zero. $\alpha$ equals the height of $x^t$. The term $\alpha - D^V_i(x^t, x^u) > 0$ and $\alpha$ will make sure the attraction force existing, which will be explained in Sect. 4.1. $|V^r_{i,x}(x^t)|$ is the absolute of relative velocity of $x^t$ and $x^u$. We adopt linear combination to model the final attraction force:

\[
F^V_{\text{att}}(x^t, x^u) = \lambda F^V_{\text{att}}(x^t, x^u) + (1 - \lambda)F^V_{\text{att}}(x^t, x^u)
\]

(5)

$\lambda$ is the tuning weight. We set $\lambda = 0.5$ in our experiment.

2.2 Analysis for Grouping

Sometimes, inter-person occlusion is observed without any attraction because of the camera settings (such as the angle from the camera), e.g., when pedestrians walk together. Thus, we employ grouping to aid in the solving of this problem.

Grouping, people intend to walk together, is another kind of social interaction that occurs frequently in the tracking scenario [16]. Because this work performs tracking online, we cannot access the complete pedestrians’ trajectories to derive global groups, as in offline tracking methods [7], [17]. In this work, we instead infer a simple frame-wise pairing grouping relation.

Persons in the same group tend to walk at similar speeds and in similar directions. We still consider the persons $x^t$ and $x^u$ for example, and the group of $x^t$ is defined as:

\[
\mathcal{G}^{ij} = 1, \text{ if } \left\{ \begin{array}{l} \Delta D^V(x^t, x^i) = 0 \; ; \mathcal{G}^{ij} = 0, \text{ if else } \end{array} \right.
\]

(6)

where $\Delta D^V(x^t, x^i)$ and $\Delta D^V(x^t, x^i)$ are the average changes in the distance between $x^t$ and $x^i$ for the last few successive frames (in our experiments, we used three). An evolution grouping $n_c \times n_c$ matrix $G_M$ is constructed in this way. $n_c$ is the number of pedestrians in group $G^n$. The element in $G_M$ is $G^{ij}$, $j = 1, 2, 3, \ldots$

The matrix initially contains all zeros, and it will be updated after each frame. $G^{ij}$ ($j = 1, 2, 3, \ldots$) consider the information from several previous frames. Such accumulative inference can provide a robust estimation of the grouping.

For a valid grouping pair $x^a$ and $x^b$, if DPM (Deformable Part based Model detector) [18] fails to match the tracker of $x^b$ because of occlusion, it will be assumed that the output velocity of $x^b$ is equivalent to $x^a$.

3. Occlusion Handling Strategy

3.1 Hierarchical Tracking-by-Detection

The online tracking-by-detection approach is one of the most commonly used methods for multi-object tracking. It combines discriminative [19] and generative methods [20]. These methods treat frame-by-frame data association as a pair-wise assignment problem, which matches the detection results with tracking. In our work, we adopted the hierarchical data-association method. Assume that in frame $t$, all of the detection inputs are considered to be from detection division $DE$, and the tracking results are considered to be from target division $TR$. Candidate $CA$ is the subset of $DE$, which is used to represent new objects that appear in the scene. To sum up, we have $DE + TR$ as the input for every frame. To prevent tracker associates with false positive detection candidates, we follow the procedure suggested by [13], the tracker will be generated when one candidate is matched for at least two consecutive frames. Without associating detection results, the tracker would only survive for limited number of frames (2 in our case). This is for checking if new detection results would match the tracker.

To assign the correct detection to the correct tracking result, we used one matching score by computing the likelihood between detections and tracking results. The matching score (M) includes several components. In our case, we use $M = Pos \cdot Size \cdot App$, where we define $Pos \cdot Size$ as the overlap between the detector (de) and tracker (tr): $\frac{de \cap tr}{de \cup tr}$.

For the appearance, we employ the Hellinger distance of the HSV color histogram. It involves computing the histogram of both the detector and tracker on the HSV color space. In order to deal with cases such as when there is varying illumination and occlusion, we keep the color histogram information of the first frame and last frame that the object has correctly tracked. After obtaining those matching scores, the Hungarian algorithm [21], which consider all the possible match and will provide the best results by efficiently global search, provides the best match. The HSV color-feature selection in our experiments is also able to achieve a fair result compared with other methods.

For each object, the tracking result and its matched detection are outputted, if there is one; otherwise, only the tracking result is used as the output.
in the experiments. Only the pairs meet such threshold will be further associate by multiplying appearance feature. If attraction exists by our deeply spatial-temporal analyzing, we consider the occlusion is caused by attraction force. As we have already pointed in previous section, attraction force based occlusion rationale is to predict occlusion for $t + 1$. Therefore the data-association performs in $t + 1$ as well. Assuming that $d_m, tr_n$, where $m, n$ is the index of detection results and tracking results, are within the same occlusion group, we propose to obtain optimal matching score $\hat{M}$ by dynamic programming search for

$$\hat{M} = \arg\max_S \sum_{(m, n) \in S} \sum_{k=1}^{k-1} \left( M(d_m, tr_n) - \gamma F_{att}^k(x^k, x^n) \right),$$ \hspace{1cm} (8)$$

with $(k = 1, 2, 3 \ldots)$. \(S\) represents all possible assignments in this occlusion group adhering to the 1-to-1 mapping constraint $F_{att}^k(x^k, x^n)$ is attraction force between $x^k$, $(k = 1, 2, 3 \ldots)$ and $x^n$ (the tracker $tr_n$ could provide this information) in this occlusion group, and $k$ is the index for targets involved in this occlusion group. This attraction force is utilized to penalize the occlusion. \(\gamma\) is weight for $F_{att}^k$. We fix $\gamma = 0.1$ in our experiments.

4. Experiments

4.1 Initialization of Social Interaction

In terms of practical issues, it is necessary to define where the attraction would happen w.r.t. each target. As noted in [22], the modeling for any two objects in the scene is unnecessary and meaningless. Thus, the search region for each object is set to eliminate the objects that are too far away to have an attraction. We manually set the search region as a square shape to eliminate the objects that are too far away to have an attraction. The size of one side of the search region is considered to be twice the height of the object. The Euclidean distance between the center point of the bounding boxes of objects is employed to estimate the distance between objects. Only the objects are within the search region of other objects, and without any occlusion are initialized for attraction forces. In addition, if the overlapping area of two bounding boxes exceeds 40%, we consider the spatial information to be invalid in order to avoid potential errors. Furthermore, the example that to deal with size of the object may lead to a loss of information. For instance, $x^1$ and $x^2$ are objects with larger and smaller sizes, respectively. When we consider the attraction of $x^3$, in addition to all of the objects within the search region, we need to estimate the attraction of $x^1$. If we determine the attraction between $x^1$ and $x^2$ for $x^1$, this information is stored and considered for $x^2$.

It is worth noting that, the spatial temporal relation between targets may vary through time. Strictly constrain the distance difference in Eq. (2) and Eq. (6) to zeros in not practical. Therefore, in our experiments, the distance difference would be checked every frame for attraction, and the average distance difference for three frames for grouping.
4.2 Datasets

To better evaluate the capability of our method, we tested our approach on the widely used PETS2009 S2L1, S2L2 dataset, and TUD-Crossing dataset. Most of the challenges involving these two datasets are the frequent occlusions caused by the dynamic motion of pedestrians. To obtain a fair comparison, we used the groundtruth of the PETS dataset and the TUD-Crossing groundtruth. All of the persons present in the scene have been annotated.

4.3 Evaluation Metrics

To measure the performance, we adopted the CLEAR MOT metrics. The metrics include: 1) Multiple Object Tracking Accuracy (MOTA, a larger value is better), which returns an accuracy score; 2) Multiple Object Tracking Precision (MOTP, a smaller value is better), which considers the intersection union of bounding boxes; 3) Mostly Tracker (MT); and 4) Mostly lost (ML). We adopted the procedure proposed by Breitenstein et al. [21], where the results are re-evaluated using the 2D spatial coordinates to decide the position of the groundtruth of the targets.

Because our aim is to investigate the advantage of our occlusion-handling method, we chose the following state-of-the-art methods for comparison. The classic online tracker [13] with a naive occlusion reasoning, a tracking framework employing similar hierarchical data association [24] without occlusion handling, the method utilizing scene knowledge [4] to tackle the occlusion issue, and the most recent approach [3]. As an example, a visual comparison between our method and Breitenstein et al. [13] in two datasets illustrate the advantage of our proposed method.

Figure 2(a) depicts our results on the PETS 2009 S2L1 and S2L2 datasets compared with the state-of-the-art method. As expected, when there is inter-person occlusion (the last two frames in each row), the proposed method is better than that of Breitenstein et al. [13]. The attraction force between the yellow, blue, and red persons leads to occlusion (the second frame in the first row), and our approach is capable of associating the detector with the correct tracker. Moreover, the results demonstrate that the proposed method can handle severe occlusion (the last frame). Note that the red and blue pedestrians are paired at the beginning of this example (shown in the first frame).

Table 1 compares the performance of the tracker on the PETS2009 S2L1 and S2L2 datasets. Utilizing the scene knowledge, the camera calibration makes the approach of [4] outperform the other methods; however, our approach performs favorably compared to most of the online trackers. Our novel occlusion-handling method improves the capability to deal with the occlusion of the tracker. Our occlusion-handling method makes the tracker more robust than the method that employs a similar hierarchical data-association scheme [24]. Thus, we achieved a better MOTA score and a significantly improved MOTP score.

Considering the more challenging TUD-Crossing sequences for inter-person occlusion, the quantitative results are shown in Table 1. In this experiment, we choose to use the method proposed by Wu et al. [3] instead of that by Possegger et al. [4]. Our method provides better scores in both MOTA and MOTP compared to other approaches. The visual results compared to those by Breitenstein et al. [13] are shown in Fig. 2(b). Our framework is boosted by the DPM model, which is capable of distinguishing that persons are very close, and it therefore gives much better detection results when compared to [13] (where the arrow points). In this experiment, the green and blue persons are considered as a group. Our proposed method deals with occlusion better than [13] (the red person and the group). In contrast, the state-of-the-art method fails to distinguish the two pedestrians, while the tracker suffers when the targets undergo occlusion.

The results obtained for both datasets confirm that our proposed method is beneficial when employing the occlusion-handling method. Spatial-temporal information provides a different perspective, which can be more reliable when combined with the appearance information by our way. Furthermore, the experimental results suggest that the use of scene knowledge [4] may lead to further improvements in the tracking performance.

4.4 Results and Analysis

The results on all datasets confirm that our method are beneficial by employing occlusion handling method. Spatial-temporal information provides a different perspective, which can be more reliable when combining with ap-

---


Table 1: Comparison of different online tracking methods on PETS 2009 dataset and TUD-Crossing dataset (down). We list the results of state-of-the-art directly from published literature.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>MOTA (%)</th>
<th>MOTP (%)</th>
<th>MT (%)</th>
<th>ML (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PETS</td>
<td>Ours</td>
<td>94.1</td>
<td>75.2</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Breitenstein et al. [13]</td>
<td>79.7</td>
<td>56.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Possegger et al. [4]</td>
<td>98.1</td>
<td>80.5</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. [24]</td>
<td>93.4</td>
<td>68.2</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td>PETS</td>
<td>Ours</td>
<td>68.4</td>
<td>60.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Breitenstein et al. [13]</td>
<td>50.0</td>
<td>56.3</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Possegger et al. [4]</td>
<td>66.0</td>
<td>64.8</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. [24]</td>
<td>66.7</td>
<td>58.6</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>TUD-Crossing</td>
<td>Ours</td>
<td>92.2</td>
<td>80.2</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Breitenstein et al. [13]</td>
<td>84.3</td>
<td>71.0</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Wu et al. [3]</td>
<td>90.6</td>
<td>76.9</td>
<td>100.0</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td>Zhang et al. [24]</td>
<td>71.3</td>
<td>67.5</td>
<td>100.0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
4.5 Run Time Performance

The unoptimized code of entire algorithm is implemented with MATLAB. All results are obtained on a single machine with 3.2 GHz CPU and 8G memory. The average speed is about 7 fps for all datasets. This performance also outperforms [13] (2 fps), and on par with [4], [24] (7 fps). To the best of our knowledge, the major consuming time of our algorithm attributes to the DPM detector. On the other hand, [24] simply consider the hierarchical data association without considering the impact of social interaction, which significantly reduces the computation cost. Surprisingly, our performance could be on par with [4], which is offline method.

5. Conclusion and Future Work

In this study, we proposed a novel inter-person occlusion-handling method based on attraction forces and grouping. We performed an extensive analysis of every possible attraction scenario, and performed occlusion handling at the data-association level. In addition, we used grouping to assist us to solve occlusion issues. The experimental results obtained show that our method is comparable with, or even better than the state-of-the-art method.

In future work, we aim to further explore the spatial-temporal information. We intend to apply the proposed method to crowd-scene analysis because a spatial-temporal perspective provides useful information. Another ongoing study involves the use of action recognition to add more information. This is expected to be successful because action in videos is usually considered as one source of spatial-temporal information.

References

Yuke Li is currently pursuing his Ph.D. in State Key Lab of LIESMARS, Wuhan University. His research interests focus mainly on multi-targets tracking, objects detection and crowd behavior analysis.

Weiming Shen is a Professor at State Key Lab of LIESMARS, Wuhan University. His research interests include image processing and computer vision. He has successfully accomplished several national projects as project leader.