What Do I See? Modeling Human Visual Perception for Multi-person Tracking*

Xu Yan, Ioannis A. Kakadiaris, and Shishir K. Shah

Department of Computer Science, University of Houston
Houston, TX 77204-3010, USA
{xyan5,ioannisk}@uh.edu, sshah@central.uh.edu

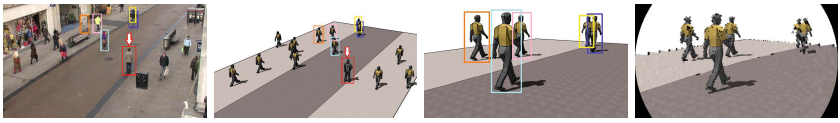
Abstract. This paper presents a novel approach for multi-person tracking utilizing a model motivated by the human vision system. The model predicts human motion based on modeling of perceived information. An attention map is designed to mimic human reasoning that integrates both spatial and temporal information. The spatial component addresses human attention allocation to different areas in a scene and is represented using a retinal mapping based on the log-polar transformation while the temporal component denotes the human attention allocation to subjects with different motion velocity and is modeled as a static-dynamic attention map. With the static-dynamic attention map and retinal mapping, attention driven motion of the tracked target is estimated with a center-surround search mechanism. This perception based motion model is integrated into a data association tracking framework with appearance and motion features. The proposed algorithm tracks a large number of subjects in complex scenes and the evaluation on public datasets show promising improvements over state-of-the-art methods.

1 Introduction

Multi-person tracking is a fundamental problem for many computer vision tasks, such as video surveillance and activity recognition. The computer vision community has begun to explore social behavior modeling to improve accuracy of multi-target tracking systems in recent years. Various social behavior models [29,24,39,31] have been explored and incorporated into the multi-person tracking frameworks. Unlike the traditional motion model, the social behavior model, in essence, treats human motion as the result of both a person's intention and their interaction with environment rather than the outcome of a motion dynamics model alone. This is a critical aspect of tracking humans and enables incorporation of the basic understanding that human beings invariably will make motion decision based on their intent and understanding of the environment. In general, typical social behavior models are built on constraints over spatial proximity and

* This work was supported in part by the US Department of Justice, grant number 2009-MU-MU-K004. Any opinions, findings, conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of our sponsors.

D. Fleet et al. (Eds.): ECCV 2014, Part II, LNCS 8690, pp. 314–329, 2014.



(a) (b) (c) (d)

Fig. 1. Examples of reconstructed virtual world. (a) The original surveillance image, (b) the virtual vision image, (c) the first-person view image from the person under the arrow, (d) the retinal mapping image. The bounding boxes in (a,b,c) with same color represent the same person.

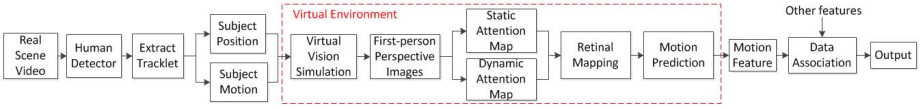


Fig. 2. System Framework. The components in red outline are implemented in virtual environment.

treat nearby subjects and objects with equal importance [38,2,18]. However, a person does not plan his/her movements based on a holistic understanding of the scene but reasons about it based on the local field of visual perception [17]. Therefore, in this paper, we propose building a perception based motion model from the first-person perspective. Intuitively, a person does not react to all subjects in his/her perspective with equal intensity. For example, a person will react strongly to a person moving faster in their direction as compared to someone moving slower. In other words, a person moving quickly towards one will take priority in one's perception and hence in their motion planning. We argue that people's attention has two kinds of variations: (1) spatial variations that are related to subjects that are near or far; and (2) temporal variations that are related to subjects that are moving fast or slow. To explore a more realistic motion model, we propose an attentive vision based tracking framework.

Benfold and Reid [6] utilized a person's head pose to locate areas of attention to guide surveillance systems. However, this information was not incorporated into a multi-person tracking framework. In our case, to visualize the scene from each person's point of view we utilize the ***virtual vision simulation*** [36] so that the scene can be rendered graphically and further used to simulate a first-person view assuming the camera to be located at the head height for each person in the scene. Figure 1(a) shows the real world, (b) shows the virtual scene, and (c) shows the first-person view image of person in the red bounding box in (a) and (b). Finally, Figure 1(d) shows the retinal mapping of the first-person view image of the specific person based on the log-polar transformation wherein the center of the first-person view image is assumed to be the focal point. Alternate approaches for simulating the scene can also be utilized [32].

We generate "attention maps" of the simulated first-person view image that guides the person's motion as shown in Figure 2. The static attention map is built based on human detection, which treats human subjects in the first-person

view image as obstacles. In this paper, we assume that the human motion is dominated by the intent of obstacle avoidance. The dynamic attention map is derived from optical flow displacement of human subjects in a person's view. Human subjects further away or moving slowly will have a smaller optical flow displacements than those in closer proximity or moving fast. Further, the optical flow displacement from first-person view image, when mapped according to retinal mapping, implicitly incorporates the effect of motion direction in which humans subjects moving towards the person along the direction of the person's focal point will exhibit expansion and occupy more area than those moving away from the person. After combining static and dynamic attention maps, retinal mapping is overlaid on the combined map to mimic human retinal vision, i.e., spatial regions far from individual's visual center will have low attention and hence lower spatial resolution and vice-versa for closer regions. The final attention map combines spatial and temporal variations of the scene as per the person's visual priority. Our method identifies regions of high interest from subject's attention map that guides the estimation of subject's next movement and serves as a novel feature in a person tracking framework. The advantage of visual attention over direct use of motion information is that it provides a reasonable mechanism to estimate the motion probability while automatically weighting the proximal and peripheral information together. The key contributions of our work are as follows:

- *Perception based multi-person tracking.* We simulated the virtual vision and get the first-person view image. Such transformation facilitates intuitive analysis of human perception and reaction to subjects in the environment and induces a more realistic motion model. This also serves to enhance social behavior models by weighting relationship graphs.
- *Attentive vision model.* We propose an attentive vision model that approximates the spatial and temporal variance of human attention. The combined attention map enables motion path prediction of a person without explicit knowledge of other person's motion. Our model predicts human motion and is combined with data association for tracking.

We define human motion as a direct consequence of human attentive vision system. The problem is then transformed into a human attentive vision modeling problem (Sec. 3), which operates in a virtual simulation world that has the same physical world coordinates as the real world. We show how to integrate attention features into a tracking-by-detection framework (Sec. 4). Finally, we test our approach in real world challenging surveillance videos and evaluate the tracking performance in comparison to other tracking methods (Sec. 5).

2 Related Work

Multi-person Tracking. Tracking-by-detection has becomes increasingly popular for multi-person tracking due to the improvement of human detector. Progress on tracking-by-detection can be attributed to development in two areas,

both of which bring the benefit to tracking performance. The first is the design of efficient data association methods. Brender *et al.* [10] used maximum weighted independent set to converge to a data association optimum. Andriyenko *et al.* [4] combined discrete with continuous optimization to solve both data association and trajectory estimation. Butt *et al.* [13] use Lagrangian relaxation to transfer the global data association to solvable min-cost flow problem. The second area is in the learning of discriminative features. In this category, multi-person tracking algorithms either exploit appearance variance feature [16,9] or model complex motion dynamics feature [33,26]. We contribute to build discriminative motion feature in this work.

Social behavior modeling has attracted more attention with its ability to quantify complex human interactions. Luber *et al.* [24] proposed to use repulsion effects to incorporate scene obstacles. Choi *et al.* [14] considered the group motion dynamics within a joint prediction model. Yan *et al.* [39] integrated the social attraction and repulsion effects into an interactive tracking framework. Qin *et al.* [31] and Bazzani *et al.* [5] exploited the social group effect associated with the tracking performance. Manocha *et al.* [8,23] leveraged reciprocal velocity obstacles model to take into account local interactions as well as physical and personal constraints. All the aforementioned works treat the social behavior from surveillance camera view angle instead of understanding social behavior from subject's own viewpoint. In this paper, we model the target motion behavior from the first-person view and utilize it for multiple target tracking. To the best of our knowledge, no previous tracking method has leveraged first-person perspective.

Visual Attention Modeling. By mimicking the human vision system, computational visual attention modeling is investigated by psychologists, microbiologists, and computer scientists. A number of computational models of attention are proposed and can be categorized based on whether they are biological, purely computational, or hybrid [15]. All plausible biological methods are directly or indirectly inspired by cognitive concepts. In contrast, Ma *et al.* [25] proposed a method based on local contrast for generating saliency maps that is not based on any biological model. Achanta *et al.* [1] had incorporated both biological and computational parts in their method. Our work falls in the area of purely computational methods. Related work in crowd simulation [27,21] has leveraged human visual attention to model the motion of virtual agents in a synthesized environment.

3 Attentive Vision Modeling

Given a configuration $C^t = \{c_i^t\}$ of subjects ($i = 1 \dots N$) at time t, each subject is modeled as $c_i^t = (p_i^t, s_i^t, a_i^t)$, where p_i^t denotes the world coordinate position, s_i^t its speed, and a_i^t its motion angle. Our method models the human perception of each subject i at the time step t based on the configuration C^t. For simplicity, we will explain one subject's attentive vision model in a scene with a fixed number of subjects. This can easily be generalized to an arbitrary number

of subjects. Unlike previous approaches, we don't assume each person's prior knowledge about other subjects' position.

3.1 Virtual Vision Simulation

We assume a person's consistent moving direction in the next step $t+1$ is same as the person's current motion direction. Based on the calibration of the real scene (Figure 1(a)) and the output of human detection, we can get the position parameter p_i^t for each subject i. The motion parameters s^i and a^i estimation will be explained in section 4. Here we assume we have the parameters c_i^t for each subject. To simplify the configuration, we also set every person's height as 1.7 meter and the eye position is 1.6 meter from the ground, which is also set as the first-person view camera's position. Using the configuration C^t, we construct the virtual scene as shown in Figure 1(b) with virtual vision simulator [36]. In the simulator, we simulate human motion based on the start point, end point and the time we set to match the estimated speed. All movements are assumed to be piece-wise linear. In the virtual scene, the first-person view image is generated by putting the virtual camera at the virtual person's head location and directed towards the virtual person's moving direction in the simulated world. The focal length is fixed for each person. Here we assume the head pose is same as the subject motion direction. An example of a first-person view image is shown in Figure 1(c). The first-person view image shares the same world coordinate with virtual vision image and real world image. In the following sections, all computations of attentive vision are performed on first-person view images. The corresponding retinal mapping image is shown in Figure 1(d) further explained in the following section.

3.2 Attention Map

Visual saliency is one of the most popular computational model for visual attention [19]. Similar to saliency based attention model [28], we compute an attention map that leverages both static and dynamic components of attention. The attention map is built as shown in Figure 2 (red outline). The first step is to construct static and dynamic maps, then to overlap retinal mapping on the combined map.

Static Map. With virtual scene, all the pedestrian's motion are simulated with virtual agents that have the same velocity as the real world scene. The images of first-person perspective are collected from virtual vision simulator for frame $\{1, \ldots, i, \ldots, K\}$. Background subtraction is performed to detect the human subjects within the controlled foreground-background contrast in virtual scene [30]. The static map is built based on human detection results in frame 1. The output of human detection of frame 1 is denoted as $R^1 = \{r_1^1, \ldots, r_n^1\}$ where r_n^1 is represented by binary foreground mask. The static map of human attention is modeled as $S_s = r_1^1 \cup \ldots \cup r_n^1$.

(a) (b) (c)

Fig. 3. The static map in first-person perspective view. (a) The over-head view image.(b) The first-person view image. (c) The static map is generated based on the human detection.

Dynamic Map. Human perception is sensitive to moving subjects and human attentive vision treats moving subjects with different velocities differently. The dynamic map is built to address the temporal variance component of attentive vision. Optical flow (O_x^i, O_y^i) is calculated for frame $\{2, \ldots, i, \ldots, K\}$, which implicitly models the relative motion between observer's and all the other subjects' motion [11]. With the virtual vision images, the human in $\{2, \ldots, i, \ldots, K\}$ frames is detected by background subtraction and the locations are denoted as $R^{2,\ldots,i,\ldots,K}$. We set $K = 25$ in this paper. The motion saliency in frame i is defined as

$$M^i(x,y) = \begin{cases} sqrt((O_x^i)^2 + (O_y^i)^2) & (x,y) \in R^i \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

The final dynamic map combines all the motion saliency denoted as $S_d(x,y) = max\{M^2(x,y), \ldots, M^i(x,y), \ldots, M^K(x,y)\}$, which is determined by taking the maximum of motion intensity. A dynamic map example for one person is shown in Figure 4(a).

Static-Dynamic Map Combination. We hypothesize that the human perception drives attention to specific areas when the motion intensity in that region is above a certain threshold. Thus the combination of static and dynamic map is fulfilled in a motion-conditioned strategy. The combined attention map is computed as follows:

$$S(x,y) = \begin{cases} 1 & \text{if } S_d(x,y) \geq \epsilon \text{ or } S_s(x,y) = 1 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where, ϵ denotes the threshold on motion intensity and is set to 0.1 in this paper. After combination, a binary mask is generated and is overlaid on the original image as shown in Figure 4(b) and 4(c). A crucial point to note here is that even though subjects receive higher perceptual attention, the regions they occupy may have lower probability as potential future target positions.

Retinal Mapping. Attentive vision refers to the reaction of people according to the visual stimuli in a dynamically changing environment, which is characterized by selective sensing in space and time as well as selective processing with respect to a specific task [34]. Selection in space involves the splitting of the visual field

(a) (b) (c)

Fig. 4. (a) The dynamic map is generated based on virtual vision simulation. (b) The combined static-dynamic attention map. (c) The combined map mask is applied on first-person perspective image.

in a high resolution area, the fovea, and a space-variant resolution area, the periphery, which are denoted as retinal mapping. Log-polar transformation is the most common method to represent visual information with a space-variant resolution [37] and to achieve retinal mapping. The log-polar transformation conserves high resolution in the center of the image and the resolution gradually decreases away from center.

We denote (x, y) for the image coordinates and $(r_{(x,y)}, \theta_{(x,y)})$ for the corresponding polar coordinates and r_{max} denotes the maximum value of $r_{(x,y)}$. The polar mapping of image pixel (x, y) with origin (x_0, y_0) is defined as

$$r_{(x,y)} = \sqrt{(x - x_0)^2 + (y - y_0)^2}, \text{ and } \theta_{(x,y)} = tan^{-1}(\frac{y - y_0}{x - x_0}). \tag{3}$$

The foveal region is defined as a round disk with the radius r_0 and origin (x_0, y_0). The image in the foveal region retains uniform resolution while the non-foveal region exhibits decreasing resolution, which is also used to indicate the importance of observations . We apply the log-polar transformation on the non-foveal part of a first-person perspective image, which is defined as the ring-shaped area $r_{max} > r_{(x,y)} > r_0$. The unified retina mapping is defined as:

$$r'_{(x,y)} = \begin{cases} r_{(x,y)} & r_{(x,y)} < r_0 \\ log(r_{(x,y)}) & r_{max} > r_{(x,y)} > r_0 \end{cases} \tag{4}$$

and $\theta'(x, y) = \theta(x, y)$. With the transformed log-polar coordinates, the quantization is applied along θ' and r' axes that results in G and R elements, respectively. As shown in Figure 5, each pixel (x, y) undergoes a transform to the log-polar space and the log-polar space is quantized. The retinal mapping of combined static-dynamic attention map is computed based on the remapping of log-polar space that transforms the log-polar image back to the Cartesian space. The remapping follows the Eq. 3 and 4 utilizing the inverse mapping of θ' and r' to x' and y', respectively. Certain number of pixels will be allocated as the same intensity value due to the quantization in log-polar space. After doing so, we get the retinal mapping on combined attention map as shown in Figure 5(b). Another attention search map is generated for motion prediction as shown in Figure 5(c). For attention search map, we compute the mean of the mapped

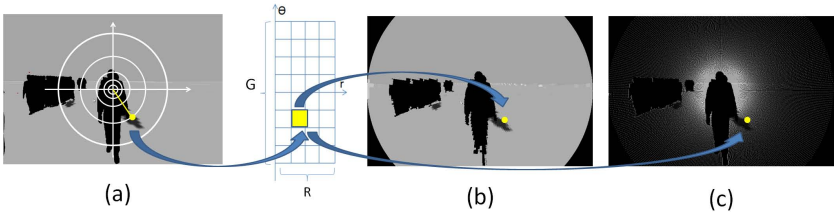


Fig. 5. The diagram of retinal mapping. (a)The first-person view image overlaid by static-dynamic map. (b) Retina mapping image. (c) Attention search map.

pixel locations and assign the intensity value from the log-polar space to the pixel position nearest to the computed mean position. The remaining pixels are assigned a value of zero. This allows us to generate a sparse map where the pixels that do not have a value of zero represent positions that can be probable locations for a target's next position.

3.3 Motion Prediction Based on Attentive Vision

This paper assumes that people follow their intuition, which means that people will find the most feasible and most attentive point as their destination. We divide this process into two step. The first step is to find the most attentive

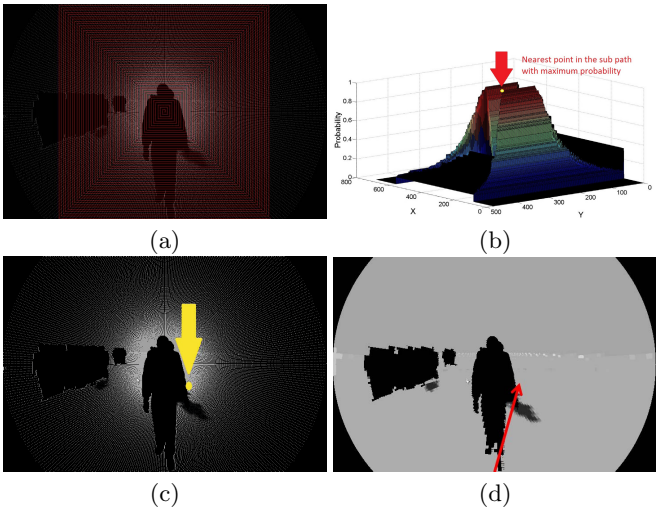


Fig. 6. (a) Center-surround search path. Red line is a sub-path, which is sparse here for visualization purpose. (b) The generated 3d probability map. The yellow point represents the nearest point in the sub path with maximum probability of being the destination point. X and Y are the original image coordinates and Z is the probability. (c) Potential destination point in first-perspective view image. (d) The calculated moving angle based on attentive vision.

sub-path based on attention search map (Figure 6(a)). A sub-path is defined as a line between two consecutive corners in the center-surround path as shown with red color in Figure 6(a). The probability of each sub-path in attention map is denoted as

$$P_{path} = \frac{m_{valid}}{m_{total}} \tag{5}$$

where, m_{valid} is the number of pixels that are not equal to zero in the attention map along the sub-path and m_{total} is the total number of pixels in the sub-path. Following the center-surround search path, the probability map of attentive vision is generated as shown in Figure 6(b). The sub-paths with maximum probability are selected as most attentive sub-path by exhaustive search.

For the second step, we calculate the corresponding world coordinate of each pixel in previous optimal sub-paths. With known observer's position, the point with the shortest distance to the observer is selected as potential destination from the optimal sub-path as shown in Figure 6(c). The predicted human motion direction π^{att} is calculated correspondingly based on the vector from the current position to found destination and is depicted in Figure 6(d). This is used to guide tracking later due to the shared world coordinate between the observer and the surveillance camera's view.

4 Tracking Framework

To reduce the computation load and for more accurate subject motion estimation, we leverage a two-stage tracking framework. In first stage, we extracts basic tracklets $\{T_1, \ldots, T_i, \ldots, T_N\}$ for each subject i in which $T_i = \{c_i^{t_i^b}, \ldots, c_i^{t_i^e}\}$ and t_i^b and t_i^e denote the begin and end time frame of T_i. The motion parameters s_i^t and a_i^t are estimated from basic tracklets. With these parameters, we simulate the virtual vision as shown in section 3 and get the motion prediction with attentive vision modeling. In second stage, we combine the predicted motion feature and other features and accomplish the tracklets association.

In first stage, we leverage common method to extract basic tracklet based on position, size and color histogram similarity in consecutive frames [31]. The color similarity constraint is also applied between current frame and first frame of tracklet. The detail of second stage is further explained in section 4.2 and 4.3.

4.1 Tracklet Association Formulation

We transform the tracklet association as 2D linear assignment problem on a bipartite graph. Given a set of tracklets $\mathbb{T} = \{T_1, T_2, \ldots, T_N\}$, we define a pairwise cost matrix H, in which h_{ij} denotes the cost that tracklet j is linked as first tracklet after tracklet i. The data association is formulated as

$$\arg\min_{\{i,j\}} \sum_{i=1}^{N} \sum_{j=1}^{N} h_{ij} x_{ij} \quad s.t. \begin{cases} \sum_{j=1}^{N} x_{i,j} = 1; \\ \sum_{i=1}^{N} x_{i,j} = 1; \\ x_{ij} \in \{0, 1\} \end{cases} \tag{6}$$

where $x_{ij} = 1$ iff tracklet j immediately follows tracklet i, otherwise, $x_{ij} = 0$. The cost is defined as the combination of five features including our attentive vision feature:

$$h_{ij} = \beta \cdot \Phi(T_i, T_j) \cdot Z(\Delta t) \tag{7}$$

where, $\beta = [\beta_1; \beta_2; \beta_3; \beta_4]$ is a vector of model parameters and set empirically in this paper, $\Phi(\cdot) = [\phi_1(\cdot), \phi_2(\cdot), \phi_3(\cdot), \phi_4(\cdot)]$ represents the association feature set, and $Z(\cdot)$ is the time gap component defined by an exponential model:

$$Z(\Delta t) = \begin{cases} \lambda^{\Delta t - 1} & 1 \leq \Delta t \leq \xi \\ \infty & \Delta t < 1 \text{ or } \Delta t > \xi \end{cases} \tag{8}$$

where ξ is the threshold of time gap and $\Delta t = t_j^b - t_i^e$.

4.2 Features Extraction

Given each tracklet pair (T_i, T_j), four features are calculated to get the association cost. The color feature ϕ_1 is build based on the 3D color histogram in the Red-Green-Intensity (RGI) space with 8 bins per channel. We perform a kernel density estimate for both the tracklets across their live frames. The similarity between two kernels $g(T_i)$ and $g(T_j)$ is measured by the Bhattacharyya coefficient B given by:

$$\phi_1 \propto \exp(-B[g(T_i), g(T_j)]). \tag{9}$$

The speed feature ϕ_2 is modeled by the Normal distribution: $\phi_2 \propto \mathcal{N}(\mu_j^s; \mu_i^s, \sigma_i^s)$ where, $\mu_j^s = mean(\sum_{t=t_j^b}^{t_j^e} s_j^t)$ is the average speed of T_j in its living period and μ_i^s, σ_i^s is the mean and variance of T_i's speed.

The angular likelihood is divided to two angular regions. The first one incorporates the attentive vision feature that assumes the next tracklet should appear at the predicted angle. It is modeled by the *von Mises* distribution [35], which is formulated as:

$$\phi_3 = \frac{e^{\kappa cos(\pi - \pi^{att})}}{2\pi I_0(\kappa)}\ , \tag{10}$$

where $I_0(.)$ is the modified Bessel function of order zero, and π denotes the motion angle between the spatial location of the middle point of tracklet i and the corresponding location of T_j. The π^{att} is our attentive vision model's predicted angle. κ corresponds to variance in a normal distribution and is set empirically. To get the informative attentive vision feature, the human motion direction history should be estimated accurately. Due to the uncertainty of detection output, we design a threshold strategy to estimate the human motion direction. When the basic tracklet is shorter than 10 frames, we compute the average optical flow to estimate the motion direction and we rule out the region overlapped by other tracklets. Otherwise, the motion direction is computed based on tracklet position information.

The second angular feature models smooth motion and penalizes motion change. This is described by the normal distribution; $\phi_4 \propto \mathcal{N}(\mu_j^a; \mu_i^a, \sigma_i^a)$, where

μ_j^a is the moving angle mean of T_j, μ_i^a and σ_i^a are the moving angle mean and variance of T_i.

4.3 Data Association

Given the cost matrix H, we solve the assignment problem through a strategy similar to the cut-while-linking strategy proposed in [31]. The cost matrix H is extended to H^{new} to solve the initialization and termination of tracks, which is defined as,

$$H^{new} = \left[\begin{array}{cccc|cccc} h_{11} & h_{12} & \dots & h_{1N} & \tau & \infty & \dots & \infty \\ h_{21} & h_{22} & \dots & h_{2N} & \infty & \tau & \dots & \infty \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ h_{n1} & h_{n2} & \dots & h_{NN} & \infty & \infty & \dots & \tau \\ \hline \infty & \infty & \dots & \infty & \infty & \infty & \dots & \infty \\ \infty & \infty & \dots & \infty & \infty & \infty & \dots & \infty \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \infty & \infty & \dots & \infty & \infty & \infty & \dots & \infty \end{array}\right]. \tag{11}$$

The thresholds τ decides when a trajectory ends and is fixed for each scene. When h_{ij} exceeds τ, the link between two tracklets is cut and the track will be linked to extended columns which indicates the track terminates. The initialization of tracks is solved along with determined termination. The extended version of data association formulation is defined as

$$\arg\min_{\{i,j\}} \sum_{i=1}^{2N}\sum_{j=1}^{2N} h_{ij}^{new} x_{ij} \quad s.t. \begin{cases} \sum_{j=1}^{2N} x_{i,j} = 1; \\ \sum_{i=1}^{2N} x_{i,j} = 1; \\ x_{ij} \in \{0,1\} \end{cases} \tag{12}$$

The optimal association is solved by Munkres' assignment algorithm [12].

5 Experiments

We evaluate how attentive vision helps to improve multi-person tracking on two public datasets: TUD stadtmitte [3] and TownCentre [7]. We follow the popular evaluation metrics [22], which includes mostly tracked trajectories (MT), mostly lost trajectories (ML), fragments (Frag) and ID switches (IDS). In addition, we also report the false positive rate (FPR) of our results on each dataset. The TUD statmitte dataset has a short video, but with very low camera angle and frequent full occlusions among pedestrians. The TownCentre video is a high definition video with 1920×1280 resolution. This sequence is very crowded with frequent occlusion and interaction among pedestrians. The pedestrians appearing briefly at the image boundaries are excluded. We also collected a video in an outdoor uncontrolled environment. It is a high definition video with 1280×720 resolution and 1200 frames in total. This sequence is crowded with 40 trajectories in total. The activity inside is challenging for tracking algorithms since a large amount of interactions are observed among the people. Walking, skateboarding and biking activity also exists in the scene. We have manually annotated the video to identify the locations and provide unique IDs for all the people in the video.

5.1 Component-Wise Evaluation

To understand the benefit of the attentive vision feature proposed in this paper, we first present the component-wise evaluation. The baseline method turns off the attentive vision feature and re-tuning to the best performance while the default methods keep all the merits of the proposed method. Table 1 presents results of quantitative comparison. The default method out-performs the baseline method in most measures across all the datasets.

Table 1. Component-wise evaluation on each dataset. The best result is in bold.

Dataset	With attentive vision	MT	ML	Frag	IDS
TUD stadtmitte	No	60.0%	**0.0**%	3	2
TUD stadtmitte	Yes	**70.0**%	**0.0**%	**2**	**1**
TownCentre	No	81.3%	6.2%	**33**	45
TownCentre	Yes	**85.6**%	**4.8**%	43	**19**
OURS	No	47.5%	20.0%	22	21
OURS	Yes	**77.5**%	**10.0**%	**13**	**18**

5.2 Comparative Evaluation

To compare fairly with different tracking method, we use the same detector's output. For TUD stadtmitte, we use the same detection and groundtruth provided by [40] and show comparable performance. The quantitative results are show in Table 2. We can see that our result is comparable or better than state-of-the-art methods. Our result is better than *Energy Min* [3], *Disc-Continue* [4] and *PRIMPT* [20] as our attentive vision incorporated model gives more informed prediction. Our approach does not provide an obvious advantage over *Online CRF* [40] since this video has low camera angle and several very short tracklets, which makes it difficult to estimate the tracklet motion direction. In this case, the power of online learned appearance model in *Online CRF* gives more benefit than motion prediction. Some sample tracking results are shown in Figure 7(a). The FPR of our method is 3.2%.

For TownCentre dataset, we use the original detection and groundtruth provided by [7], which are used in [31], and we show improvement by incorporating the attentive vision features. The quantitative comparison is shown in Table 2. The results show that the attentive vision based tracking model outperforms *Basic affinity model* [31] and *SGB model* [31] in terms of MT, ML, and IDS. Fragment of trajectories under our model increased due to threshold setting in cut-while-linking strategy. Example qualitative result is shown in Figure 7(b,c,d). The FPR of our method is 7.6%.

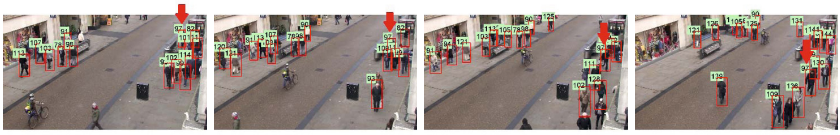
We compare our method's performance with *SGB model*. We also replace the attentive vision model in our framework with *LTA model* [29] and keep all the other components fixed. The quantitative results are shown in Table 2, which show that attentive vision model outperforms *SGB model* and *LTA model* in terms of MT, ML and Frag. LTA does a little better in IDS than our model. The qualitative evaluation is shown in Figure 7(e). The FPR of our method is 5.9%.

Frame 22 Frame 63 Frame 107 Frame 165

(a) Tracker under heavy occlusion and interaction: Object 1 is tracked correctly.

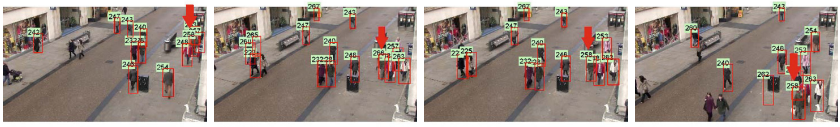
Frame 252 Frame 418 Frame 539 Frame 635

(b) Long-term tracking under full occlusion, abrupt motion change and miss detection: Object 26 is tracked correctly in spite of significant change of motion direction.



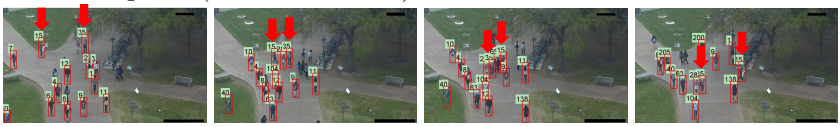
Frame 1287 Frame 1353 Frame 1418 Frame 1509

(c) Robust tracking in densely populated regions: Object 97 change the motion paths frequently due to the oncoming crowd.



Frame 2620 Frame 2697 Frame 2703 Frame 2777

(d) ID fragment correction: Object 258 suffers from ID fragment (but not ID switch) which is corrected in Frame 2703.



Frame 53 Frame 141 Frame 201 Frame 295

(e) Attention vision prediction: Object 15 distracted from large amount of moving subjects which is corrected predicted by attentive vision modeling and recovered in Frame 201 and Frame 295.

Fig. 7. Tracking results of our approach on TUD statmitte, TownCentre and our campus datasets. For visualization purpose, certain false positive trajectories are not shown.

Table 2. Comparison of results on TUD statmitte, TownCentre and Our dataset. The best result is in bold. [31] (a) and [31] (b) represent the baseline method and proposed method in [31] respectively.

Dataset	Method	MT	ML	Frag	IDS
TUD stadtmitte	Andriyenko *et al.* [3]	60.0%	**0.0**%	4	7
TUD stadtmitte	Kuo *et al.* [20]	60.0%	10.0%	**0**	1
TUD stadtmitte	Andriyenko *et al.* [4]	60.0%	**0.0**%	1	4
TUD stadtmitte	Yang *et al.* [40]	**70.0**%	**0.0**%	1	**0**
TUD stadtmitte	Proposed method	**70.0**%	**0.0**%	2	1
TownCentre	Qin *et al.* [31] (a)	76.8%	7.7%	37	60
TownCentre	Qin *et al.* [31] (b)	83.2%	5.9%	**28**	39
TownCentre	Proposed method	**85.6**%	**4.8**%	43	**19**
OURS	Qin *et al.* [31]	45.0%	22.5%	24	22
OURS	Pellegrini *et al.* [29]	62.5%	15.0%	19	**16**
OURS	Proposed method	**77.5**%	**10.0**%	**13**	18

6 Conclusion

We have presented a novel tracking method using an attentive vision model where motion analysis is performed in the first-person view. The attentive vision is created from virtually reconstructed scene. A visual attention map is generated based on attentive vision mechanism, including both static and dynamic components. The most feasible path taken by the person is searched and decided from this constructed map. The predicted motion direction is integrated into data-association tracking with color and motion features. The association is solved by a greedy algorithm. As the experiments show, the proposed approach achieves promising improvements on different public datasets. Finally, the performance of the algorithm could be improved if we enhance the short tracklet motion estimation method and the virtual simulation details.

References

1. Achanta, R., Susstrunk, S.: Saliency detection for content-aware image resizing. In: Proc. ICIP, pp. 1005–1008 (2009)
2. Ali, S., Shah, M.: Floor fields for tracking in high density crowd scenes. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 1–14. Springer, Heidelberg (2008)
3. Andriyenko, A., Schindler, K.: Multi-target tracking by continuous energy minimization. In: Proc. CVPR, pp. 1265–1272 (2011)
4. Andriyenko, A., Schindler, K., Roth, S.: Discrete-continuous optimization for multi-target tracking. In: Proc. CVPR, pp. 1926–1933 (2012)
5. Bazzani, L., Cristani, M., Murino, V.: Decentralized particle filter for joint individual-group tracking. In: Proc. CVPR, pp. 1886–1893 (2012)
6. Benfold, B., Reid, I.: Guiding visual surveillance by tracking human attention. In: Proc. BMVC, pp. 1–11 (2009)

7. Benfold, B., Reid, I.: Stable multi-target tracking in real-time surveillance video. In: Proc. CVPR, pp. 3547–3464 (2011)
8. Bera, A., Manocha, D.: Realtime multilevel crowd tracking using reciprocal velocity obstacles. CoRR abs/1402.2826 (2014)
9. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: Proc. ICCV, pp. 1515–1522 (2009)
10. Brendel, W., Amer, M., Todorovic, S.: Multiobject tracking as maximum weight independent set. In: Proc. CVPR, pp. 1273–1280 (2011)
11. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE T-PAMI 33(3), 500–513 (2011)
12. Burkard, R., Dell'Amico, M., Martello, S.: Assignment Problems. Society for Industrial and Applied Mathematics, Philadelphia (2009)
13. Butt, A.A., Collins, R.T.: Multi-target tracking by lagrangian relaxation to min-cost network flow. In: Proc. CVPR, pp. 1846–1853 (2013)
14. Choi, W., Savarese, S.: Multiple target tracking in world coordinate with single, minimally calibrated camera. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 553–567. Springer, Heidelberg (2010)
15. Filipe, S., Alexandre, L.A.: From the human visual system to the computational models of visual attention: a survey. Artificial Intelligence Review 39(1), 1–47 (2013)
16. Grabner, H., Bischof, H.: On-line boosting and vision. In: Proc. CVPR, pp. 260–267 (2006)
17. Hari, R., Kujala, M.V.: Brain basis of human social interaction: From concepts ot brain imaging. Physiological Reviews 89(2), 453–479 (2009)
18. Kim, S., Guy, S.J., Liu, W., Lau, R.W.H., Lin, M.C., Manocha, D.: Predicting pedestrian trajectories using velocity-space reasoning. In: Proc. WAFR, pp. 609–623 (2012)
19. Koch, C., Ullman, S.: Shifts in slective visual attention: Towards the underlying neural circuitry. Human Neurbiology 4, 219–227 (1985)
20. Kuo, C., Nevatia, R.: How does person identity recognition help multi-person tracking? In: CVPR, pp. 1217–1224 (2011)
21. Lee, K.H., Choi, M.G., Hong, Q., Lee, J.: Group behavior from video: A data-driven approach to crowd simulation. In: Proc. SCA, pp. 109–118 (2007)
22. Li, Y., Huang, C., Nevatia, R.: Learning to associate: Hybridboosted multi-target tracker for crowded scene. In: Proc. CVPR, pp. 2953–2960 (2009)
23. Liu, W., Chan, A.B., Lau, R.W.H., Manocha, D.: Leveraging long-term predictions and online-learning in agent-based multiple person tracking. CoRR abs/1402.2016 (2014)
24. Luber, M., Stork, J., Tipaldi, G., Arras, K.: People tracking with human motion prediction from social forces. In: Proc. ICRA, pp. 464–469 (2010)
25. Ma, Y., Zhang, H.: Contrast-based image attention analysis by using fuzzy growing. In: Proc. International Conference on Multimedia, pp. 374–281 (2003)
26. Mei, X., Ling, H.: Robust visual tracking using l_1 minimization. In: Proc. ICCV, pp. 1436–1443 (2009)
27. Ondrej, J., Pettré, J., Olivier, A.H., Donikian, S.: A Synthetic-Vision Based Steering Approach for Crowd Simulation. In: Proc. SIGGRAPH, pp. 123:1–123:9 (2010)
28. Ouerhani, N.: Visual attention: from bio-inspired modeling to real-time implementation. Ph.D. thesis, Univeristy of Neuchâtel, Switzerland (2003)
29. Pellegrini, S., Ess, A.: K.Schindler, van Gool, L.: You'll never walk alone: modeling social behavior for multi-target tracking. In: Proc. ICCV, pp. 261–268 (2009)

30. Piccardi, M.: Background subtraction techniques: a review. In: Proc. IEEE conference on Systems, Man and Cybernetics, pp. 3099–3104 (2004)
31. Qin, Z., Shelton, C.R.: Improving multi-target tracking via social grouping. In: Proc. CVPR, pp. 1972–1978 (2012)
32. Qureshi, F., Terzopoulos, D.: Smart camera networks in virtual reality. In: Proc. International Conference on Distributed Smart Cameras, pp. 87–94 (2007)
33. Ross, D.A., Lim, J., Lin, R., Yang, M.: Incremental learning for robust visual tracking. IJCV 77(1), 125–141 (2008)
34. Schwartz, E.L., Greve, D.N., Bonmassar, G.: Space-variant active vision: Definition, overview and examples. Neural Networks 8(7), 1297–1308 (1995)
35. Song, B., Jeng, T.-Y., Staudt, E., Roy-Chowdhury, A.K.: A stochastic graph evolution framework for robust multi-target tracking. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part I. LNCS, vol. 6311, pp. 605–619. Springer, Heidelberg (2010)
36. Thiebaux, M., Marshall, A., Marsella, S., Kallman, M.: Smartbody: Behavior realization for embodied conversational agents. In: Proc. AAMAS, pp. 1151–1158 (2008)
37. Traver, V.J., Bernardino, A.: A review of log-polar imaging for visual perception in robotics. Robotics and Autonomous Systems 58(4), 378–398 (2010)
38. Treuille, A., Cooper, S., Popović, Z.: Continuum crowds. In: Proc. SIGGRAPH, pp. 1160–1168 (2006)
39. Yan, X., Kakadiaris, I., Shah, S.: Predicting social interactions for visual tracking. In: Proc. BMVC, pp. 102.1–102.11 (2011)
40. Yang, B., Nevatia, R.: An online learned CRF model for multi-target tracking. In: Proc. CVPR, pp. 2034–2041 (2012)