Figure-Aware Tracking under Occlusion from Monocular Videos

Xue Wang, Qing Wang
School of Computer Science
Northwestern Polytechnical University
Xi'an, R.P. China
xwang@mail.nwpu.edu.cn, qwang@nwpu.edu.cn

***Abstract*—In this paper, we propose a figure-aware tracking framework incorporating figure/ground repulsive forces in a simultaneous detectlet classification and clustering problem in the joint space of detectlets and trajectlets for monocular videos. Without depth/disparity, fine-grained trajectlets tend to cause under-segmentation of similarly moving objects or over-segmentation of articulated objects into rigid parts. Detectlets represented by the bounding boxes only help avoiding under-segmentation of similarly moving objects under canonical pose, while do no good for improving the over-segmentation problem. Pose estimation, though not accurate, is often sufficient to segment human torso from its backgrounds and induce figure/ground repulsions, which could reduce the risk of both under-segmentation and over-segmentation. Figure-aware mediation encodes repulsive segmentation information in trajectory affinities and provides more reliable model aware information for detectlet classification. Our algorithm can track objects through sparse, inaccurate detections, persistent partial occlusions, deformations and background clutter.**

Keywords- Multiple object tracking; Video segmentation; Normalized cuts; Figure/Ground segmentation; Pose estimation

I. INTRODUCTION

We address the problem of multiple object segmentation and tracking in crowded scenes with only monocular information available. We propose a framework incorporating figure/ground repulsions induced by pose estimation in a mediated detectlet classification and clustering formulation.

Frameworks combining bottom-up and top-down information have been widely used in segmentation and recognizing static images [1, 2], pose estimation [3, 4] and video segmentation and tracking [5, 6]. For video segmentation and tracking, such approaches usually employ dense spatio-temporal representations of trajectories or super-pixel clusters that adapt to motion discontinuities across object boundaries. However they often have difficulty handling deformable or articulated motion or close object interactions, resulting in under-segmentation of similarly moving objects or over-segmentation of articulated objects into rigid parts. The state-of-the-art object tracking algorithms [7, 8] link detection over time, which suffer from detection under persistent partial occlusion, widely deformed poses and cluttered background.

Combining different grained tracking units is capable of varying opportunistically between a whole object tracker during full visibility of the template and fine-grained point trackers during partial occlusions [6]. Dense trajectories help to link two detectlets together that do not overlap in time but belong to the same identity or propagate object location for the missing detection. Detectlets represented by the bounding boxes only help avoiding under-segmentation of similarly moving objects under canonical pose, while do no good for improving the over-segmentation problem. Either adopting more grouping information for better trajectlet affinity computation or using a post-processing for cluster combination would be a direction. In the paper, we focus on the trajectlet affinities. Apart from motion and spatial affinities, depth/disparity provides good cue for reducing the risk of under-segmentation or over-segmentation over the depth discontinuities. When depth/disparity is not available, such as with monocular videos, the figure/ground repulsive segmentation information can be encoded into trajectlet affinities and provides more reliable model aware cues to detectlet classification.

The key insight underlying this work is to attempt to incorporate figure/ground repulsions induced by pose estimation in the mediated segmentation graph (Section 3). Pose estimation, though it often fails for lower limbs, is accurate for torso of people [9]. Such reliable detections help segmentation of body pose by adjusting motion affinities to conform to the figure/ground segmentation of detected human body [4]. To evaluate the performance of our algorithm we compare against a traditional globally-optimal tracking algorithm [7] and alternative versions of our figure-aware tracking framework on several challenging datasets, which is described in Section 4.

II. RELATED WORKS

Researchers have explored various ways of linking sparse detections in time, such as using region information [10] or body part tracking [11]. The authors of [7] propose a global solution by computing the shortest path on a flow network to track a variable number of objects. In [12], a sparsity-induced tracklets association algorithm is used to deal with complete occlusion. Since detections are often sparse in time and spatially inaccurate, fine-grained trajectories which are dense in both space and time, could provide complementary information for tracking and be adapt to the changing visibility mask of occluded objects. A two-granularity tracking algorithm mediating group cues from detectlets and point trajectlets is proposed in [6],

978-1-4799-6854-1/14 $31.00 © 2014 IEEE
DOI 10.1109/ICVRV.2014.13

which is cast as simultaneous detectlet classification and clustering in the joint space of tracklets and trajectories. Their algorithm relies greatly on the acquisition of depth information of the scenes.

Figure/ground segmentation is a well-studied but still challenging problem in the computer vision community. An accurate figure/ground decision is based on many cues, such as 2D visual features (brightness, color, texture gradients etc.) [25], motion [26, 27], depth [28] and model-related physical constraints (e.g. kinematic constraints for articulated objects) [29, 30]. Moreover, figure/ground segmentation can also be a good cue for other higher level visual task, such as video segmentation [22] and tracking [23, 24]. Fragkiadaki et al. [4] use motion grouping affinities and on lower limbs and figure/ground repulsions from shoulder detections to help the detection of highly deformable body poses. Improved body pose estimations are further used to benefit motion estimation of lower limbs. Optical flow information is also exploited in pose estimation problem as a cue either for body part detection or for pose propagation from frame-to-frame [13]. Brox et al. [14] introduce a pose tracking system that interleaves between contour-driven pose estimation and optical flow pose propagation from frame to frame.

III. FIGURE-AWARE MEDIATION

We use figure/ground segmentation to help multi-object tracking from monocular videos. With stereo videos, disparity maps can be computed and used for fine-grained trajectlets clustering and detection. When stereo information is not available, however, trajectory affinities tend to leak across objects with similar motion, over-segment articulated objects into rigid parts, or fail to delineate stationary objects from the background. Additionally the bounding box representation of detectlets suffers from cluttered background. By mediating the figure/ground repulsions induced by pose estimation to the affinity matrix, we could suppress affinities between foreground and background trajectlets and boost affinities between both foreground or background trajectlets, resulting in more accurate clusters.

A. Two-granularity Tracking

Trajectlets and detectlets are obtained following the frame optical flow field [15, 16] and by conservatively linking detections between consecutive frames respectively. Stylized body poses are covered by an abundance of training examples in current vision datasets [20] and can be reliably detected with state-of-the-art detectors [9].

Given a video frame I_t of video sequence I, we define $\mathbf{T} = \{tr_a, \mathrm{a} = 1 \cdots n_T\}$ to be the set of point trajectories, where tr_a is a sequence of space-time points $tr_a = \{(x_a^t, y_b^t), t \in T_a\}$ and T_a is the frame span of tr_a. For each pair of trajectories (tr_a, tr_b), we compute affinity $\mathbf{A}_T(tr_a, tr_b)$ which encodes their long range spatial and motion similarity [17].

Let $\mathbf{D} = \{dl_p, p = 1 \cdots n_D\}$ denote the set of detectlets, where dl_p represents a sequence of bounding boxes for each person hypothesis. We set $dl_p = \{(\mathrm{box}_p^t, f_p^t), t \in T_p\}$, where box_p^t is the detection bounding box at frame I_t, f_p^t is the corresponding detection score and T_p is the frame span of the detectlet. We define the confidence of detectlet dl_p to be the sum of confidences of its detection responses $\mathrm{f}_p = \sum_{t \in T_p} f_p^t$. We set affinities $\mathbf{A}_D(dl_p, dl_q)$ between detectlets dl_p and dl_q that do not overlap in time according to the anchoring score between their closest in time detections [6].

The associations $\mathbf{C}(tr_a, dl_p)$ between trajectlets tr_a and detectlet dl_p are computed according to their spatio-temporal overlap:

$$\mathbf{C}(tr_a, dl_p) = 1 \ \mathbf{if} \ \forall t \in T_a \cap T_p, (x_a^t, y_a^t) \in \mathrm{box}_p^t . \quad (1)$$

Thus these relationships can be summarized in a $n \times n$ affinity matrix $\mathbf{A}$:

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_T & \mathbf{C} \\ \mathbf{C}^{\mathrm{T}} & \mathbf{A}_D \end{bmatrix} \begin{matrix} \} \ n_T \\ \} \ n_D \end{matrix}, \quad (2)$$

where $n = n_T + n_D$, n_T is the number of trajectlets and n_D the number of detectlets.

B. Induced Affinity Graph $\mathbf{A}(f)$

Considering the inaccurate detection responses in cluttered scenes, in order to exclude the false associations between trajectlet-detectlet pairs and the affinity contradictions between trajectlet-trajectlet pairs, a model-aware affinity graph $\mathbf{A}(h^s)$ is introduced in [6], which is cast as simultaneous detection tracklet classification $h^s \in \{0,1\}^{n_D \times 1}$ and detectlet-trajectlet co-clustering problem. Only selected detectlets dl_p, $h^s(p) = 1$ can claim trajectlets through associations in $\mathbf{C}(h^s)$ and induce model-aware affinity graph $\mathbf{A}_T(h^s)$.

Let $\mathbf{R}_t = \{r_i^t, i = 1 \cdots n_R^t\}$ denote the set of image regions of frame I_t. We use r_i to refer to both the region and its corresponding pixel set. Each detection p_m in the poselet detection set $\mathbf{P} = \{p_m, m = 1 \cdots n_p\}$ implicitly induces figure/ground repulsive forces between the regions associated with its interior and exterior. Repulsions $\mathbf{R}(r_i, r_j \mid \mathbf{P})$ are induced between foreground and background regions of each detector response. Let $z_m^F, z_m^B \in \{0,1\}^{n_R \times 1}$ denote foreground and background region indicators and let M_m and U_m denote the pixel set overlapping with p_m and the pixel set outside a contour of displacement Δ to figure/ground outlier respectively. We

have:

$$z_m^F(i)=\delta(\frac{|r_i\cap M_m|}{|r_i|}>0.9),\ i=1\cdots n_R, m=1\cdots n_D$$
$$z_m^B(i)=\delta(\frac{|r_i\cap U_m|}{|r_i|}>0.5),\ i=1\cdots n_R, m=1\cdots n_D, \quad (3)$$

where δ is the Dirac delta function being 1 if its argument is true and 0 otherwise. Repulsions $\mathbf{R}(r_i,r_j\,|\,\mathbf{P})$ are induced between foreground and background regions of each detector response:

$$\mathbf{R}(r_i,r_j\,|\,\mathbf{P})=\max_{m|d_m\in\mathbf{P}} z_m^F(i)z_m^B(j)+z_m^B(i)z_m^F(j). \quad (4)$$

Induced trajectlet affinities take the final form:

$$\mathbf{A}_T(f)=\mathbf{A}_T(h^s,tr_a,tr_b)\cdot\lambda\max_{t\in T_a\cap T_b}(1-\mathbf{R}(r_i^t,r_j^t\,|\,\mathbf{P})), \quad (5)$$

where $(x_a^t,y_a^t)\in r_i^t$, $(x_b^t,y_b^t)\in r_j^t$, $\mathrm{dis}_{a,b}$ denotes the maximum Euclidean distance between tr_a and tr_b , $\mathrm{vel}_{a,b}$ denote the maximum velocity difference during their time overlap and λ is a scalar parameter boosting affinities between those regions belonging to both foreground or background. Mediated trajectlet affinities help avoid over-segmenting articulated objects into rigid parts (Fig. 1). We summarize the figure-aware relationships in the following combined affinity graph $A(f)$:

$$\mathbf{A}(f)=\begin{bmatrix}\mathbf{A}_T(f) & \mathbf{C}(h^s)\\ \mathbf{C}^{\mathrm{T}}(h^s) & \mathbf{A}_D\end{bmatrix}\begin{matrix}\}\ n_T\\ \}\ n_D\end{matrix}. \quad (6)$$

C. Optimization

We formulate figure-aware tracking as the simultaneous detectlets classification and clustering in the joint (selected) detectlet and trajectlet space. Let $X_k\in\{0,1\}^{n_T\times 1}$ and $Y_k\in\{0,1\}^{n_D\times 1}$ denote indicator vectors for trajectlets and detectlets clusters respectively, where $k=1\cdots K$ and K being the total number of clusters. We have the following joint optimization over detectlet classification h^s and co-clustering (X,Y) :

$$\max_{h^s,X,Y,K}\sum_{k=1}^{K}\mathrm{ncut}(\mathbf{A}(f),X_k,Y_k)\cdot\mathrm{confidence}(Y_k)$$
$$s.t.\quad \sum_{k=1}^{K}X_k=1_{n_T},\ \sum_{k=1}^{K}Y_k=h^s,\ \mathrm{align}(X_k,\mathrm{Y}_k)>th. \quad (7)$$

Figure 1. Figure/ground repulsions help avoiding over-segmenting articulated objects into rigid parts. Left: different optical flows of different body parts in the cyan bounding box. Middle: shape mask aligned to edges. Right: trajectlet cluster labeled in cyan covers almost the whole body.

The cost function of Eq. 8, which is similar to [6], requires the detectlet/trajectlet cluster (X_k,Y_k) to be a salient (stable) group under the normalized cut criterion [21]. We sample h^s according to detectlet confidence f . By varying the number of segments K and the minimum trajectlet length we compute multiple segmentations in the induced affinity graph $\mathbf{A}(f)$, resulting in a pool of detectlet-trajectlet clusters. We further prune clusters with low alignment score $\mathrm{align}(X_k,Y_k)$ .We obtain the tracking solution by sequentially choosing the best scoring cluster from the remaining ones, that does not overlap with already chosen ones.

IV. Experiments

The proposed tracking framework is evaluated in CAVIAR dataset [1] and TUD Stadtmitte dataset [18]. Ground-truth is both provided as a set of pedestrian boxes which we link manually into ground-truth tracklets. We further evaluate our framework in NWPU-Kids, an indoor multi-view children social activities dataset containing 15 sequences. Each sequence features several synchronized video streams filming, either from egocentric viewpoints taken by GoPro HERO3 cameras or conventional third person viewpoints taken by Microsoft Kinect cameras. Ground-truth is provided in the form of segmentation masks for all visible targets every six frames (approximately 0.2 seconds) in each sequence. We only use Kinect RGB videos for multi-object tracking evaluation in the paper. Both CAVIAR and TUD Stadtmitte are pedestrian detection and tracking datasets with small body deformation. While in NWPU-Kids widely deformed poses and cluttered background are very challenging.

To provide a basis of comparisons for our figure-aware mediation, an ablative analysis of our system is performed, by comparing to the following baselines: 1) mediation without figure/ground assignment information, which is similar to [6], 2) mediation by co-embedding, which clusters in the unmodified affinity matrix $\mathbf{A}$ of Eq.3 and 3) bottom-up trajectlet clustering in $\mathbf{A}_T$. We additionally compare our method to a data association algorithm using network flow [7], abbreviated as GlobalDA.

CLEAR MOT metrics [19] is used for performance evaluation. We compute a one-to-one assignment of hypotheses to ground-truth objects in each frame, measuring intersection over union of boxes on CAVIAR and TUD Stadtmitte and of segmentation masks on NWPU-Kids. We report numbers of miss detections, false alarms, id-switches and tracking accuracy in Table 1

The numerous miss detections of bottom-up clustering are due to stationary objects as well as to pedestrian groups with similar motion. Figure-unaware mediation is sensitive to noisy, false alarm or fast movement of lower limbs. Detectlet sparsity is verified in the large number of detectlet

[1] http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/

miss detections, resulting in an even worse miss detection of GlobalDA. Our figure-aware mediation framework outperforms, being robust to false alarm detectlets (cluttered background) and over-segmentation of articulated objects into rigid parts (fast movement of lower limbs). Since trajectlet is discontinuous under complete occlusion, ID-switch of our method and trajectlet-based methods are often higher than GlobalDA which employs a greedy but globally optimal algorithm for solving the problem of multiple object tracking. It is worth noting that our method reaches an accuracy of 73.1% on NWPU-Kids, which is 10.5% higher than figure-unaware mediating and better than the performance on the other two datasets. The higher improvement lies in the fact that the ground-truth on NWPU-Kids is provided in the form of segmentation masks for all visible targets. We incorporate figure-aware information in the trajectlet affinities. This prevents over-segmentation of articulated objects into rigid parts and hence could output more accurate fine-grained tracking results.

Some qualitative results are displayed in Fig. 2 and Fig. 3. Figure-unaware mediation tends to be sensitive to noise, false alarm and fast movement of lower limbs. GlobalDA fails under wild pose deformation and severe occlusions. Although our method generally outperforms other methods in these datasets, it still misses some lower limbs with extreme fast movements.

TABLE 1. Tracking results on CAVIAR, TUD Stadtmitte and NWPU-Kids.

	CAVIAR				TUD Stadtmitte			
	MD(%)	FA(%)	ID-sw.	Acc.(%)	MD(%)	FA(%)	ID-sw.	Acc.(%)
Our Method	12.1	6.2	12	**82.4**	22.7	12.6	32	**73.2**
Figure-unaware Mediating	18.2	14.6	15	76.5	27.5	18.4	46	68.4
Co-embedding	29.7	30.3	22	65.2	32.0	28.7	59	54.6
Bottom-up Clustering	32.2	27.5	17	75.4	48.4	26.1	37	62.0
GlobalDA [7]	37.7	15.8	7	51.6	53.6	34.5	28	42.4
	NWPU-Kids							
	MD(%)	FA(%)	ID-sw.	Acc.(%)				
Our Method	15.6	5.4	25	**73.1**				
Figure-unaware Mediating	18.5	12.9	37	62.6				
Co-embedding	23.5	28.2	49	51.5				
Bottom-up Clustering	34.1	23.0	38	67.4				
GlobalDA [7]	58.8	12.8	22	43.0				

Figure 2. Tracking results on NWPU-Kids. Widely deformed poses and cluttered background challenges. Our method tracks targets under wildly deformable poses and partial occlusions. Figure-unaware mediation is sensitive to noisy, false alarm or fast movement of lower limbs. We can see several false alarms in the middle row. Also lower limbs, especially lower legs, are basically missed out due to the over-segmentation problem. GlobalDA fails under severe deformation or occlusions.

Figure 3. Tracking results on CAVIAR and TUD Stadtmitte. Our method tracks targets under severe partial occlusions. Figure-unaware mediation is sensitive to noisy, false alarm or fast movement of lower limbs. We can see several false alarms on both datasets. Also some lower limbs, especially lower legs, are missed out due to the over-segmentation problem on TUD Stadtmitte. While the moving is extreme fast, the performance of our method also decreases. GlobalDA fails under severe occlusions.

V. CONCLUSION

We presented a figure-aware tracking framework in a simultaneous detectlet classification and clustering problem in the joint space of detectlets and trajectlets for monocular videos. Sparse detections are used to prevent the under-segmentation of similarly moving objects under canonical pose. Figure/ground repulsions from torso detection introduced by inaccurate pose estimation are used to prevent the over-segmentation of articulated objects into rigid parts and provide more reliable model aware information for detectlet classification. Our experimental results show that figure/ground repulsive information helps getting more accurate fine-grained tracking results through sparse, inaccurate detections, persistent partial occlusions, wild deformations and background clutters.

Our method is capable of preventing the over-segmentation of articulated objects into rigid parts when the movement of lower limbs is within a reasonable range. With extreme fast movement of lower limbs, our method may fail.

In future, we will try to incorporate more model aware information in the mediation, like kinematic constraints for articulated objects.

ACKNOWLEDGMENT

We would like to thank the reviewers for helpful comments, many researchers for sharing their valuable datasets, and Kindergarten Affiliated to Northwestern Polytechnical University for assisting us completing the data collection for NWPU-Kids.

REFERENCES

[1] A. Levin, Y. Weiss, "Learning to combine bottom-up and top-down segmentation." In: ECCV. (2006)

[2] C. Pantofaru, C. Schmid, M. Hebert, "Object recognition by integrating multiple image segmentation." In: ECCV. (2008)

[3] C. Ionescu, F. Li, C. Sminchisescu, "Latent structured models for human pose estimation." In: ICCV. (2011)

[4] K. Fragkiadaki, H. Han, J. Shi, "Pose from flow and flow from pose." In: CVPR. (2013)

[5] K. Fragkiadaki, G. Zhang, J. Shi, "Video segmentation by tracking discontinuities in a trajectory embedding." In: CVPR. (2012)

[6] K. Fragkiadaki, W. Zhang, G. Zhang, J. Shi, "Two-granularity tracking: Mediating trajectory and detection graphs for tracking under occlusions". In: ECCV. (2012)

[7] H. Pirsiavash, D. Ramanan, C.C. Fowlkes, "Globally-optimal greedy algorithms for tracking a variable number of objects." In: CVPR. (2011)

[8] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L.V. Gool, "Robust tracking-by-detection using a detector confidence particle filter." In: ICCV. (2009)

[9] L. Bourdev, S. Maji, T. Brox, J. Malik, "Detecting people using mutually consistent poselet activations." In: ECCV. (2010)

[10] D. Mitzel, E. Horbert. A. Ess, B. Leibe, "Multi-person tracking with sparse detection and continuous segmentation." In: ECCV. (2010)

[11] B. Wu, R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors." In: IJCV. (2007)

[12] L. Zhang, Q. Wang, "Spatio-temporal clustering model for multi-object tracking through occlusions." In: ACCV. (2012)

[13] B. Sapp, D. Weiss, B. Taskar, "Parsing human motion with stretchable models." In: CVPR. (2011)

[14] T. Brox, B. Rosenhahn, D. Cremers, H.P. Seidel, "High accuracy optical flow serves 3-D pose tracking: exploiting contour and flow based constraints." In: ECCV. (2006)

[15] T. Brox, J. Malik, "Large displacement optical flow: Descriptor matching in variational motion estimation". In: ICCV. (2011)

[16] N. Sundaram, T. Brox, K. Keulzer, "Dense point trajectories by GPU-accelerated large displacement optical flow". In: ECCV. (2010)

[17] T. Brox, J. Malik, "Object segmentation by long term analysis of point trajectories." In: ECCV. (2010)

[18] M. Andriluka, S. Roth, B. Schiele, "Monocular 3d pose estimation and tracking by detection." In: CVPR. (2010)

[19] K. Bernardin, R. Stiefelhagen, "Evaluating multiple object tracking performance: the clear mot metrics." J. Image Video Process. (2008)

[20] M. Everingham, L.V. Gool, C.K.I. Williams, J. Winn, A. Zisserman, "The pascal visual object classes (VOC) challenge." IJCV, 88. (2010)

[21] J. Shi, J. Malik, "Normalized cuts and image segmentation." TPAMI. (2000)

[22] F. Li, T. Kim, H. Ahmad, D. Tsai, J.M. Rehg, "Video segmentation by tracking many figure-ground segments." In: ICCV. (2013)

[23] X. Ren, J. Malik, "Tracking as repeated figure/ground segmentation." In: CVPR. (2007)

[24] Z. Yin, R.T. Collins, "Shape constrained figure-ground segmentation and tracking." In: CVPR. (2009)

[25] P. Arbeláez, M. Maire, C. Fowlkes, J. Malik, "Contour detection and hierarchical image segmentation." In: PAMI. (2011)

[26] P. Sundberg, T. Brox, M. Maire, P. Arbeláez, J. Malik, "Occlusion boundary detection and figure/ground assignment from optical flow." In:CVPR. (2011)

[27] T. Brox, J. Malik, "Object segmentation by long term analysis of point trajectories." In: ECCV. (2010)

[28] L. Ladicky, J. Shi, M. Pollefeys, "Pulling things out of perspective." In:CVPR. (2014)

[29] P. Arbeláez, B. Hariharan, C. Gu, S. Gupta, L. Bourdev, J. Malik, "Semantic segmentation using inverse detectors." In:ICCV. (2011)

[30] G. Mori, X. Ren, A.A. Efros, J. Malik, "Recovering human body configurations: combining segmentation and recognition." In:CVPR. (2004)