USING MAXIMUM CONSISTENCY CONTEXT FOR MULTIPLE TARGET ASSOCIATION IN WIDE AREA TRAFFIC SCENES

Xinchu Shi[1,2]*, Peiyi Li*[2]*, Haibin Ling*[2]*, Weiming Hu*[1]*, Erik Blasch*[3]

[1]National Laboratory of Pattern Recognition, Institute of Automation, Beijing, China
[2]Department of Computer and Information Science, Temple university, Philadelphia, USA
[3]Air Force Research Lab, USA

ABSTRACT

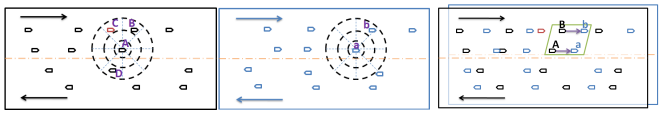
Tracking multiple vehicles in wide area traffic scenes is challenging due to high target density, severe similar target ambiguity, and low frame rate. In this paper, we propose a novel spatio-temporal context model, named *maximum consistency context* (MCC), to leverage the discriminative power and robustness in the scenario. For a candidate association, its MCC is defined as the most consistent association in its neighborhood. Such a maximum selection picks the reliable neighborhood context information while filtering out noisy distraction. We tested the proposed context modeling on multi-target tracking using three challenging wide area motion sequences. Both quantitative and qualitative results show clearly the effectiveness of MCC, in comparison with algorithms that use no context and standard spatial context respectively.

Index Terms— Context modeling, multi-target tracking.

1. INTRODUCTION

Surveillance over wide area motion imagery (WAMI) has recently been attracting increasing amount of research attention due to its wide range of applications and the advance in acquisition techniques [1–11]. However, tracking and detection of multiple moving vehicles in such scenes are challenging tasks, since a wide area traffic scene usually contains much more moving targets than traditional visual tracking scenarios do. These targets are often hard to distinguish from each other due similar appearances and small sizes in images. In addition, the low frame rate, which is typical in WAMI applications, brings extra ambiguity for associating targets over long time periods.. To address such ambiguity, researchers have investigated ways to integrate additional information, such as prior knowledge [1], scene structures [6], and target context [7, 10, 11].

Using target context to improve multiple target tracking in wide area traffic scenes is of special interest because it is relatively domain independent and therefore applicable to various situations. For a target under investigation, existing methods [7] usually model its context by the spatial distribution of other targets within its neighborhood. Such modeling implicitly assumes that the neighborhood context is reliable, which for example is true for a group of targets moving together. However, in traffic scenes it is common to see vehicles with opposite directions are spatially close to each other due to the juxtaposition of two-way lanes. Furthermore, the false positives in target detection often introduce noises into the spatial context. An example is illustrated in Fig. 1(a–b).



(a) (b) (c)

Fig. 1. Spatial context (a–b) and maximum consistency context (c). (a) Traffic scene in frame t. Circular-polar histogram centered at object 'A' represents its spatial context. (b) Traffic scene in frame $t+1$. Circular-polar diagram denotes the spatial context of object 'a'. (c) Traffics in frame t and $t+1$ are overlapped and shown in different color. Association (B,b) is the maximum consistency context of association (A,a).

Motivated by the above observation, we propose a new model named *maximum consistency context* (MCC), which is a spatio-temporal context and is robust to noises in target neighborhood. For a potential association from two consecutive frames, e.g. (**A**, **a**) in Fig. 1(c), the idea is to extract context information from only the most consistent association in (**A**, **a**)'s neighborhood, e.g. (**B**, **b**) in Fig. 1(c). With contextual information, MCC effectively reduces the disturbance from neighbor targets which either head in different directions or are false detections. Meanwhile, MCC provides strong discriminative features to guide multi-object association across frames.

2. RELATED WORK

Multiple target tracking (MTT) is a widely studied topic in computer vision (e.g. [12, 13]). The focus of this paper is on the context modeling in MTT. Contextual information plays a critical role in visual tracking. Yang et al. [14] proposed to explore the auxiliary objects which have persistent co-occurrence and consistent motion correlation with the target object, to help localize and reacquire targets. However, it is hard to mine the auxiliary objects. In Reilly's work [7],

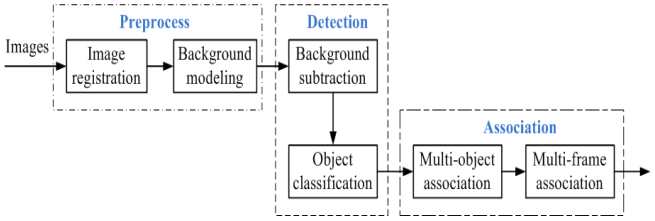


Fig. 2. The framework of wide area surveillance.

the context is the representation of geometric relationships of objects with their respective neighbors. Grabner et al. [15] proposed to use local features, which have some temporal relations with the target, to predict the target position. However, the method is computationally expensive on finding and matching features, and it is neither suitable for traffic scene. In [11], Ali et al. propose to use similar objects to predict and reacquire targets after occlusion, but they assume object detections are good enough.

Compared with aforementioned work, the main contribution is the novel context model, which fits well the task of tracking cloudy targets with mixed local motion patterns.

3. MULTIPLE VEHICLE TRACKING

3.1. Framework Overview

Following the general tracking-by-detection framework [16, 17], we treat multi-vehicle tracking as a target association problem. We roughly divide the tracking framework into three consecutive stages: preprocessing, detection, and association. The flow chart is summarized in Fig. 2. In this section we will briefly introduce the first and second stages. Then we will elaborate in detail the proposed context model and association algorithms, which are the focus of our work.

In the preprocessing, similar as in [7], we first make use of registration for image alignment. In particular, Speeded Up Robust Features (SURF) [18] features are used for the efficiency and we then fit affine models for image warping. After the alignment, background modeling is achieved through a standard median filter.

In the detection, instead of directly using the results from background subtraction, a vehicle classifier cascade is applied on the results in order to eliminate noises. The cascade is composed of two SVMs: the first SVM uses simple shape features (dimensions of a candidate target); and the second SVM uses the histogram of oriented gradient (HOG) [19] .

3.2. Multi-Object Association

Some notations are given first. We denote the image sequences as $\{I_t : t{=}1,...,n_I\}$, such that I_t is the frame at time t. The detected candidates at time t are denoted as $O^t = \{\mathbf{o}_i^t : i = 1,...,n_i\}$ containing n_i candidate targets (objects). A target $\mathbf{o}$ is defined as a vector $\mathbf{o}{=}(\mathbf{x}(\mathbf{o}),\mathbf{h}(\mathbf{o}),\theta(\mathbf{o}),a(\mathbf{o}))$ or simply $(\mathbf{x},\mathbf{h},\theta,a)$, where $\mathbf{x}$, $\mathbf{h}$, θ and a denote the location, appearance histogram, orientation and area of $\mathbf{o}$ respectively.

Without loss of generality, we assume the two frames to be associated are frames 1 and 2. To handle the missing targets and false detections in O^1 and O^2, we introduce dummy targets into the two sets and then assume they have the same number of targets, i.e., $n_1 = n_2 = n$. The multi-object association can be defined as to find the assignment $\Pi = \{\pi_{i,j}\} \in \{0,1\}^{n\times n}$ to maximize certain total association score, denoted as $\mathcal{E}(\Pi; O^1, O^2)$. The problem is formulated as

$$\max_{\Pi_n} \mathcal{E}(\Pi; O^1, O^2) = \max_{\Pi} \sum_{i=1}^{n}\sum_{j=1}^{n} \pi_{ij}(s_{ij} + c_{ij}) \,, \quad (1)$$

$$\text{s.t.} \sum_{i=1} \pi_{ij} = 1; \sum_{j=1}^{n} \pi_{ij} = 1; \pi_{ij} \in \{0,1\}; i,j \in \{1,...,n\}, \quad (2)$$

where $\pi_{ij} = 1$ (or 0) indicates there is an (or no) association between $\mathbf{o}_i^1, \mathbf{o}_j^2$; s_{ij} measures the affinity between $\mathbf{o}_i^1, \mathbf{o}_j^2$; and c_{ij} represents the context similarity between $\mathbf{o}_i^1, \mathbf{o}_j^2$.

The association without context modeling can be viewed as a special case where $c_{ij} = 0, \forall i,j$. The association problem turns to a standard integer assignment, where the Hungarian algorithm [20] provides the optimum solution.

The item s_{ij} measures the similarity between targets $\mathbf{o}_i^1 = (\mathbf{x}_i,\mathbf{h}_i,\theta_i,a_i) \in O^1$ and $\mathbf{o}_j^2 = (\mathbf{x}_j,\mathbf{h}_j,\theta_j,a_j) \in O^2$, in terms of appearance, area and orientation. Specifically, it is defined as

$$s_{ij} = \alpha s_{h,ij} + \beta s_{o,ij} + (1-\alpha-\beta)s_{a,ij}, \quad (3)$$

where s_h,s_o and s_a denote respectively for similarities in appearance, orientation and area; and α,β are weight factors.

3.2.1. *Spatial context.*

At frame t, for a target candidates $\mathbf{o}_i^t = (\mathbf{x}_i, \mathbf{h}_i, \theta_i, a_i) \in O^t$, its spatial context (SC), denoted by SC_i^t, measures the spatial distribution of other candidates in O^t. Specifically, it divides the neighborhood of $\mathbf{o}_i^t$ into $n_d \times n_o$ distance-orientation bins, and SC_i^t is then defined as a weighted $n_d \times n_o$ histogram as

$$\mathrm{SC}_i^t(p,q) = \frac{1}{Z}\sum_{\mathbf{o}_k \in \mathcal{N}_i^t} \exp\left(-(\mathbf{v}_{ik} - \mathbf{u}_{pq})^\top \Sigma^{-1} (\mathbf{v}_{ik} - \mathbf{u}_{pq})\right) \,, \quad (4)$$

where $\mathcal{N}_i^t = \{\mathbf{o}_k^t : \mathbf{o}_k^t \in O^t, \|\mathbf{x}_i - \mathbf{x}_k\|_2 \le r\}$ defines the neighborhood of $\mathbf{o}_i^t$ with radius r; $\mathbf{v}_{ik} = (\|\mathbf{x}_i - \mathbf{x}_k\|_2, \mathrm{atan2}(\mathbf{x}_k - \mathbf{x}_i))^\top$ calculates the relative distance and orientation of $\mathbf{o}_k^t$ with respect to $\mathbf{o}_i^t$; $\mathbf{u}_{pq} = (p\Delta d, q\Delta\theta)^\top$ represents the bin (p,q) such that Δd and $\Delta\theta$ are the distance and angle interval respectively; Σ is the estimated covariance matrix; and, finally, Z is the normalization constant. An illustration of spatial context is shown in Fig. 1(a–b).

With SC_i^t, c_{ij} in Eq. (1) is represented as $c_{ij} = \mathrm{sim}(\mathrm{SC}_i^1, \mathrm{SC}_j^2)$, where $\mathrm{sim}(.,.)$ defines the similarity between two histograms, which is computed using histogram intersection [21].

3.2.2. *Maximum consistency context.*

The spatial context captures rich statistics that is powerful when the target's neighborhood remains stable over time.

However, in the wide area traffic environment, this can be violated since vehicles moving on opposite directions are often close to each other. Likewise, SC can be vulnerable by inaccurately taking such noises into account. In the following, we present an alternative context modeling, which captures only the most reliable information in a target's neighborhood.

First, for two association pairs $(\mathbf{o}_i^1, \mathbf{o}_j^2)$ and $(\mathbf{o}_{i'}^1, \mathbf{o}_{j'}^2)$, we measure the consistency between them as follows,

$$\varphi\big((\mathbf{o}_i^1,\mathbf{o}_j^2),(\mathbf{o}_{i'}^1,\mathbf{o}_{j'}^2)\big)=\gamma\cos^2(\theta_{ij}-\theta_{i'j'})+(1-\gamma)\frac{2l_{ij}l_{i'j'}}{l_{ij}^2+l_{i'j'}^2}, \quad (5)$$

where $\theta_{ij} = \mathrm{atan2}(\mathbf{x}(\mathbf{o}_i^1) - \mathbf{x}(\mathbf{o}_j^2)), l_{ij} = \big\|\mathbf{x}(\mathbf{o}_i^1) - \mathbf{x}(\mathbf{o}_j^2)\big\|_2$ are the orientation and length of $(\mathbf{o}_i^1, \mathbf{o}_j^2)$ respectively, $\theta_{i'j'}, l_{i'j'}$ have similar definitions; and γ is the weight parameter.

With this definition, we define the *maximum consistency context* (MCC) for an association $(\mathbf{o}_i^1, \mathbf{o}_j^2)$ as

$$\mathrm{MCC}((\mathbf{o}_i^1,\mathbf{o}_j^2),\Pi)= \underset{(\mathbf{o}_{i'}^1,\mathbf{o}_{j'}^2)\in\mathcal{N}(\mathbf{o}_i^1,\mathbf{o}_j^2,\Pi)}{\arg\max} \varphi\big((\mathbf{o}_i^1,\mathbf{o}_j^2),(\mathbf{o}_{i'}^1,\mathbf{o}_{j'}^2)\big), \quad (6)$$

where $\mathcal{N}(\mathbf{o}_i^1, \mathbf{o}_j^2, \Pi)=\{(\mathbf{o}_{i'}^1, \mathbf{o}_{j'}^2):\mathbf{o}_{i'}^1\in\mathcal{N}_i^1, \mathbf{o}_{j'}^2\in\mathcal{N}_j^2, \pi_{i'j'}=1\}$ defines the spatio-temporal neighborhood of $(\mathbf{o}_i^1, \mathbf{o}_j^2)$, which depends on the association Π.

The proposed MCC is more flexible as it does not request the majority consistency in a target's neighborhood. In contrast, it extracts context information only from the most reliable neighbor association. Such a scheme makes MCC robust to distractions of inconsistent motions in a target's neighborhood, e.g., the vehicles running in opposite lanes in the highway scenario. It also performs robustly against false detections, as illustrated in Fig. 1.

To integrate MCC in track association, we define c_{ij} in Eq. (1) as $c_{ij} = \varphi\big((\mathbf{o}_i^1,\mathbf{o}_j^2), \mathrm{MCC}(\mathbf{o}_i^1,\mathbf{o}_j^2)\big)$. As a result, we have the following MCC-based association problem

$$\max_{\Pi}\sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{ij}(s_{ij}+\varphi((\mathbf{o}_i^1,\mathbf{o}_j^2),\mathrm{MCC}(\mathbf{o}_i^1,\mathbf{o}_j^2))). \quad (7)$$

Considering the fact that $\pi_{ij} \in \{0, 1\}$, we can rewrite (7) as

$$\max_{\Pi}\sum_{i,j=1}^{n}\pi_{ij}(s_{ij}+\max_{(\mathbf{o}_{i'}^1,\mathbf{o}_{j'}^2)\in\mathcal{N}(\mathbf{o}_i^1,\mathbf{o}_j^2)}\pi_{i'j'}\varphi((\mathbf{o}_i^1,\mathbf{o}_j^2),(\mathbf{o}_{i'}^1,\mathbf{o}_{j'}^2))). \quad (8)$$

The formulation is a quadratic integer optimization with non-smooth component (i.e., 'max'). The global optimal solution is unfortunately computationally expensive. In line with the interlocked property of this formation, we devise an iterated algorithm to optimize it, which is given in Algorithm 1.

Note that there are other approaches that can approximate the global solution for (8), such as simulated annealing and Monte Carlo methods. These solutions are however very time-consuming owing to their random property as well as the high dimensionality in our problem. Our approach, with a small number of iterations, i.e., $n_{it} = 10$, efficiently generates excellent results as shown in our experiments .

Algorithm 1 Multi-object association with MCC

Input: two consecutive candidate target sets O^1 and O^2
Output: association matrix Π
Calculate affinities $S = \{s_{ij}\}$ according to Eq. (3)
Initialization $\Pi = \mathbf{0}$;$\mathcal{E}_{max} = 0$;$C = \{c_{ij}\} = \mathbf{0}$
for $i = 1, 2, \ldots, n_{it}$ **do**
  Compute the integrated similarities $d_{ij} = s_{ij} + c_{ij}$
  Solve Eq. (9) using the Hungarian algorithm

$$\max_{\Pi}\mathcal{E}(\Pi; O^1, O^2) = \max_{\Pi}\sum_{i=1}^{n}\sum_{j=1}^{n}\pi_{ij}d_{ij}. \quad (9)$$

  Denote the solution and score by $\widehat{\Pi}$ and $\widehat{\mathcal{E}}$ respectively.
  Update C using $\widehat{\Pi}$ according to Eq. (6)
  if $\widehat{\mathcal{E}} > \mathcal{E}_{max}$ **then**
    $\mathcal{E}_{max} = \widehat{\mathcal{E}}, \Pi = \widehat{\Pi}$.
  end if
end for

3.3. Multiple frame association

We treat multi-frame tracking as a task of associating short reliable tracklets into long tracks, in a similar framework used in [22, 23]. Two procedures, reliable tracklets acquisition and tracklet-to-tracklet association, are performed iteratively.

Reliable tracklets are obtained by checking the motion smoothness of trajectories. Suppose a basic tracklet is $M_{1:t}^k = \{\mathbf{o}_k^1, \mathbf{o}_k^2, ..., \mathbf{o}_k^t\}$, it is reliable only if

$$\theta_{t-1,t}^k - \theta_{t,t+1}^k < \theta_0; \min\left\{\frac{l_{t-1,t}^k}{l_{t,t+1}^k}, \frac{l_{t,t+1}^k}{l_{t-1,t}^k}\right\} > l_0, \quad (10)$$

where $\theta_{t-1,t}^k, l_{t-1,t}^k$ are the orientation and length of association $(\mathbf{o}_k^{t-1}, \mathbf{o}_k^t)$ respectively; θ_0 and l_0 are corresponding thresholds. In this way, if association $(\mathbf{o}_k^{t-1}, \mathbf{o}_k^t)$ has inconsistency with adjoining associations, trajectory $M_{1:t}^k$ is divided into two short tracklets by breaking the association $(\mathbf{o}_k^{t-1}, \mathbf{o}_k^t)$.

Tracklet-to-tracklet association follows the similar manner as two-frame multi-target association. First, the affinity between two tracklets is constituted of appearance, motion and temporal similarities. Later, Hungarian algorithm is used here again to associate the two tracklets into the long one.

Finally, isolated detections and too short tracklets after multi-frame association are discarded as false alarms.

4. EXPERIMENT

Our experiments are conducted on the CLIF dataset [24]. CLIF is challenging with following features. 1) Large image format (4008×2672); 2) Large camera and target motion; 3)Tiny target occupancy (4∼70 pixels); 4) Similar target appearance; 5) Low frame rate sampling (2 fps); and 6) A large mount of targets (hundreds). Three sequences with 80 frames (40 seconds), 50 frames (25 seconds) and 100 frames (50 seconds) respectively are used to evaluate the proposed approach.

To study the effectiveness of the proposed context, we

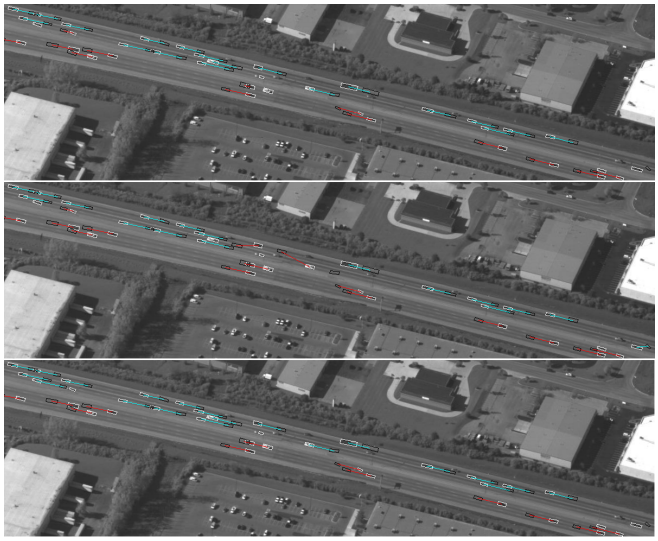


Fig. 3. Associations with different contexts on Sequence 1. Top: NoCon. Middle: SC. Bottom: MCC. Black(White) rectangle: detection in the last(current) frame. Red(blue) line: rightward(leftward) association output.

evaluate three types of associations: association without context (NoCon), with the spatial context (SC) and with the proposed maximum consistency context (MCC). Furthermore, for each method, we test two different affinity models: with appearance features and without. In all cases, Hungarian algorithm is used as the basic assignment solution.

The parameters are set as follows: γ in Eq. (5) is set as 0.5 in the first sequence and 0.8 in the other two sequences. If appearance feature is used, both α and β in Eq. (3) are set as 0.3 in all sequences. Otherwise, α and β are set as 0.5 equally. The maximum iteration number (n_{it} in Algorithm 1) is 10.

We use the precision, $100 \sum_t c(t) / \sum_t g(t)$, to measure the association performances. $c(t)$ and $g(t)$ in the precision measure denote the numbers of correct and groundtruth association respectively. Results are summarized in Table 1.

From the results, we have the following observations. First, the proposed MCC performs the best in general. Furthermore, if no appearance is used in the association, all methods have degenerated performances, yet our method is affected least. This confirms that the proposed context plays an important role in handling the motion ambiguity. Second, spatial context helps little in performance. This seemingly contradicting phenomenon can be attributed to two factors: (1) the scenes we selected contain mainly two-direction highways, which largely confuses local context as we conjectured; and (2) the vehicle detection is rather noisy, leading to unstable spatial distributions. Third, the method without using context obtains the moderate results, which attributes to kinds of well-designed affinity measures. However, when excluding the appearance information, the performance drop of NoCon is much larger than those of SC and MCC.

Qualitative results of three methods on sequence 1 are shown in Fig. 3, which is a snapshot of two-frame association.

	Seq 1		Seq 2		Seq 3	
	A+	A−	A+	A−	A+	A−
NoCon	92.1	88.5	89.5	77.8	90.7	86.0
SC	86.8	86.1	84.8	76.2	75.2	72.9
MCC	**92.5**	**90.5**	**89.6**	**80.5**	**91.3**	**89.6**

Table 1. Precisions of multi-object association. 'A+': appearance features are used; and 'A−': no appearance features are used.

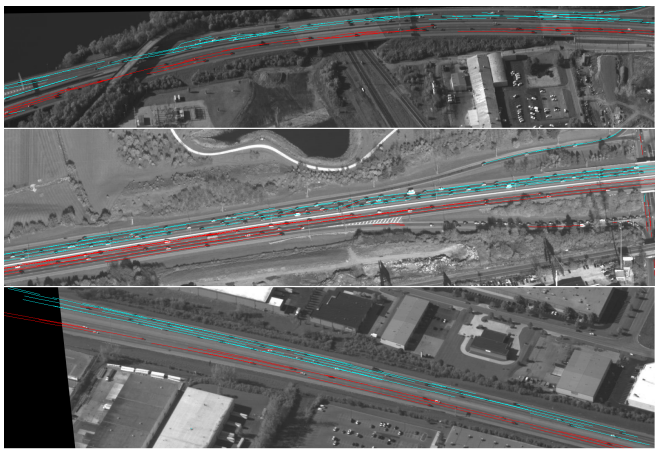


Fig. 4. Multiple object tracking results, Top: sequence 3; Middle: sequence 2; Bottom: sequence 1

Associations with SC are distracted by those cars moving in opposite directions. While NoCon is vulnerable to ambient similar objects, after embedding MCC, the motion ambiguity is greatly alleviated. With the help of neighbor association, confused object succeeds in finding true associated target, it can be seen from Fig. 3.

Results of multiple object tracking are shown in Fig. 4. It can be seen that, our approach has two merits. First, there are few wrong two-frame association. Second, our approach associates the track fragments into long tracks, which is especially useful in the case of occlusion and missing detections.

We do not have quantitative evaluation on multiple object tracking, as no groundtruth data is available. The labeling of associations in wide area traffic scenes is a very time consuming work, which we will consider in the future.

5. CONCLUSION

In this paper, we propose a novel spatial-temporal context, maximum consistency context (MCC), to assist multi-object tracking in wide area traffic scenes. By picking the most reliable ingredient in the spatial-temporal neighborhood of an association, MCC leverages the discriminative power and robustness against clutter distractions. Experiments using challenging wide area surveillance videos validate the effectiveness of the proposed approach.

Acknowledgement. This work is partly supported by NSFC (Grant No.60935002), the National 863 High-Tech RD Program of China (Grant No.2012AA012504), the Natural Science Foundation of Beijing (Grant No.4121003), and Guangdong Natural Science Foundation (Grant No.S2012020011081). Ling was supported in part by NSF Grant IIS-1218156.

6. REFERENCES

[1] S. Pellegrini, A. Ess, K. Schindler, and L. Van Gool, "You'll never walk alone: Modeling social behavior for multi-target tracking," *IEEE International Conference on Computer Vision*, pp. 261–268, 2009.

[2] J. Xiao, H. Cheng, F. Han, and H. Sawhney, "Geo-spatial aerial video processing for scene understanding and object tracking," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[3] J. Xiao, H. Cheng, H.S. Sawhney, and F. Han, "Vehicle detection and tracking in wide field-of-view aerial video," in *CVPR*, 2010.

[4] E. Blasch, G. Seetharaman, K. Palaniappan, H. Ling, and G. Chen, "Wide-area motion imagery (wami) exploitation tools for enhanced situation awareness," in *Proc. IEEE Applied Imagery Pattern Recognition (AIPR) Workshop: Computer Vision: Time for Change*, 2012.

[5] J. Prokaj, X. Zhao, and G.G. Medioni, "Tracking many vehicles in wide area aerial surveillance," in *CVPR Workshops*, 2012, pp. 37–43.

[6] J. Prokaj and G. Medioni, "Using 3d scene structure to improve tracking," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1337–1344, 2011.

[7] V. Reilly, H. Idrees, and M. Shah, "Detection and tracking of large number of targets in wide area surveillance," *European Conference on Computer Vision*, pp. 186–199, 2010.

[8] H. Ling, Y. Wu, E. Blasch, G. Chen, and L. Bai, "Evaluation of visual tracking in extremely low frame rate wide area motion imagery," in *Proc. of the Int's Conf. on Information Fusion (FUSION)*, 2011.

[9] K. Palaniappan, F. Bunyak, P. Kumar, I. Ersoy, S. Jaeger, K. Ganguli, A. Haridas, J. Fraser, R.M. Rao, and G. Seetharaman, "Efficient feature extraction and likelihood fusion for vehicle tracking in low frame rate airborne video," in *Proc. of the International Conference on Information Fusion (FUSION)*, 2010.

[10] X. Shi, H. Ling, E. Blasch, and W. Hu, "Context-driven moving vehicle detection in wide area motion imagery," in *Int'l Conf. on Pattern Recognition (ICPR)*, 2012.

[11] S. Ali, V. Reilly, and M. Shah, "Motion and appearance contexts for tracking and re-acquiring targets in aerial videos," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–6, 2007.

[12] Y. Bar-Shalom and T. Fortmann, *Tracking and data association*, Academic Press., 1988.

[13] D. Reid, "An algorithm for tracking multiple targets," *TAC*, vol. 24, no. 6, pp. 843–854, 1979.

[14] M. Yang, Y. Wu, and G. Hua, "Context-aware visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 7, pp. 1195–1209, 2009.

[15] H. Grabner, J. Matas, L. Van Gool, and P. Cattin, "Tracking the invisible: Learning where the object might be," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1285–1292, 2010.

[16] M. Andriluka, S. Roth, and B. Schiele, "People-tracking-by-detection and people-detection-by-tracking," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, 2008.

[17] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. Van Gool, "Online multiperson tracking-by-detection from a single, uncalibrated camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 9, pp. 1820–1833, 2011.

[18] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," *European Conference on Computer Vision*, pp. 404–417, 2006.

[19] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886–893, 2005.

[20] H.W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[21] M.J. Swain and D.H. Ballard, "Color indexing," *International journal of computer vision*, vol. 7, no. 1, pp. 11–32, 1991.

[22] C. Huang, B. Wu, and R. Nevatia, "Robust object tracking by hierarchical association of detection responses," pp. 788–801, 2008.

[23] C.H. Kuo, C. Huang, and R. Nevatia, "Multi-target tracking by on-line learned discriminative appearance models," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 685–692, 2010.

[24] "Afrl: Columbus large image format (clif) 2006.," https://www.sdms.afrl.af.mil/index.php?collection=clif2006.