Scene-Adaptive Hierarchical Data Association for Multiple Objects Tracking

Can Wang, *Student Member, IEEE*, Hong Liu, *Member, IEEE*, and Yuan Gao

***Abstract*—Obtaining reliable and discriminative target representation are two vital tasks for data association in multi-tracking. Pervious works always directly combine bunch of features for more discriminative target representation, but this is prone to error accumulation and unnecessary computational cost, which on the contrary may increase identity switches in data association. Moreover, reliability of a same feature in different scenes may vary a lot, especially for currently widespread network cameras, which have been settled in complex and various scenes, previous fixed feature selection scheme cannot meet general requirements. To address this problem, we propose a scene-adaptive hierarchical data association scheme, which adaptively selects features which have higher reliability on target representation in applied scene, and gradually combines features to the minimum requirements of discriminating ambiguous targets. Hierarchical feature space is constructed according to reliability of features in the multi-tracking system, and data association is conducted in different layers of the feature space adaptively. Our algorithm is validated on various challenging RGB-D and RGB datasets recorded in various indoor and outdoor scenes, for diversities of both features and scenes. Experimental results validate its effectiveness and efficiency.**

***Index Terms*—Data association, multiple objects tracking.**

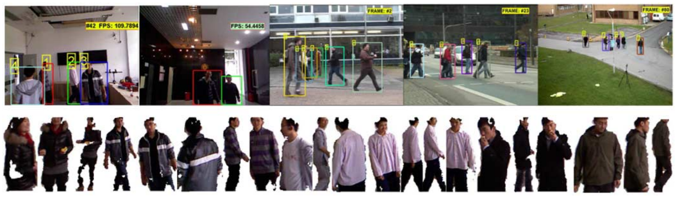

Fig. 1. Tough problems in multi-tracking. The first row shows various scene types multi-tracking applied, such as indoor, outdoor, distant, close, crowded, sparse, etc. The bottom row shows detection responses obtained from our dataset with large scale variation, frequent view-truncation, partial occlusion, and wider range of poses [10] which are all multi-tracking faces in practice.

I. Introduction

MULTI-TRACKING aims to locate moving objects, maintain their identities and retrieve their trajectories [1], in other words, to perform data association based on detection responses through a video sequence. Most valuable works have been done [2]–[5] which can be organized into two main categories: One category takes information from future frames [6]–[8] to get better association via global analysis, like global trajectory optimization [6], network flows [8], hierarchical tracklets association [9], etc. However, it is not suitable for time-critical applications and is relatively computation-consuming. The other category only considers past and current frames to make association decisions [11]–[13]. They usually relied on Kalman [14] or particle filter [15] to handle data association. Because of their recursive nature, this category is suitable for time-critical application, but it may easily lead to irrecoverable wrong data association in crowded scene with similar appearance and complicated interactions. This requires the system not only has enough ability to discriminate all targets on a given frame, but also has stable representation for each target in consecutive frames.

In order to increase the discriminating power, many pervious works usually combine a bunch of features to represent detection responses [1]–[16] and calculate the affinity matrix between them and existing tracklets. But they are with unsatisfactory performance on handling relatively challenging scenes for two reasons: First, feature representation of a same target may exhibit large variation due to illumination variation and wide range of poses. This indicates that stable representation of target is hard to obtain in challenging scenes. Second, errors of targets representation are common in cluttered scenes. For example, position of the detection response may not be exactly on the center of targets due to frequent view-truncation and partial occlusion (as shown in Fig. 1), especially in close-range scenes. This indicates same feature representation may have different reliability in different scenes and systems. Therefore, combining a bunch of features may not contribute to a better association between detection response and existing tracklets. On the contrary, features have lower reliability or discriminating power can bring adverse effect on reliable features, and also unnecessary computational cost.

In order to address above problems and to achieve a general multi-tracking algorithm suitable for network cameras in various scenes, our work focus on adaptively selecting relative reliable features in the applied scene and efficiently combining them to discriminate ambiguous targets. Our main contribution lies in two aspects: (1) We originally proposed a scene-adaptive feature selection scheme, which measures features reliability for targets representation and selects relatively reliable

Manuscript received December 19, 2013; revised March 05, 2014; accepted March 14, 2014. Date of publication March 26, 2014; date of current version April 03, 2014. This work was supported by National Natural Science Foundation of China (NSFC Grants 61340046, 60875050 and 60675025), the National High Technology Research and Development Program of China (863 Program, Grant 2006AA04Z247), the Scientific and Technical Innovation Commission of Shenzhen Municipality (Grants JCYJ20120614152234873, CXC201104210010A; JCYJ20130331144631730, and JCYJ20130331144716089), and by the Specialized Research Fund for the Doctoral Program of Higher Education (Grant 20130001110011). The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Giuseppe Scarpa.

The authors are with the Engineering Lab on Intelligent Perception for Internet of Things (ELIP) and Key Laboratory of Machine Perception (Ministry of Education), Shenzhen Graduate School, Peking University, Beijing, China (e-mail: canwang@pku.edu.cn; hongliu@pku.edu.cn; ygao@sz.pku.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LSP.2014.2313853

1070-9908 © 2014 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.

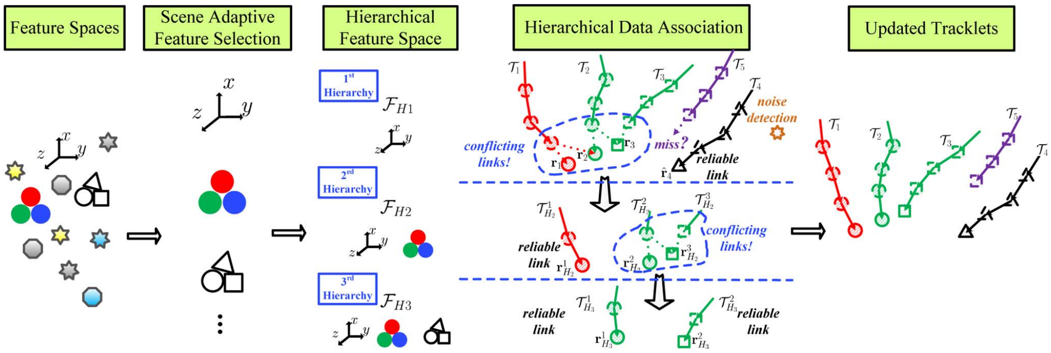

Fig. 2. A brief illustration of hierarchical data association on scene-adaptive feature space. Noise detection, miss detection, reliable links and conflicting links are unified and properly handled in this hierarchical framework.

ones for data association. This makes the algorithm much more general in the camera network with various scenes. (2) A novel hierarchical data association scheme is proposed, where features are gradually fused according to the need of discriminating ambiguous detection responses. This avoids unnecessary computation cost and reduces error accumulation compared to simultaneously fusing bunch of features.

II. Feature Selection and Data Association

A. Preliminaries

In the multi-cue data association framework, the key problem is to associate n detection responses in the current frame with m existing tracklets. Through out this paper, let $\mathcal{R}^t := \{\boldsymbol{r}_i\}_n$ denote n detection responses at frame t and let $\boldsymbol{r}_i$ denote one detection response. Let $\mathcal{T} := \{\mathcal{T}_j\}_m$ denote m existing tracklets and let $\mathcal{T}_j$ denote one tracklet, where $\mathcal{T}_j := \{\cdots, \boldsymbol{r}_j^{(t-2)}, \boldsymbol{r}_j^{(t-1)}\}$, and $\boldsymbol{r}_j^{(t)}$ denotes the detection response associated to tracklet $\mathcal{T}_j$ at time t. It can be seen from Fig. 2, one tracklet actually contains all detection responses associated to it in the past.

The most commonly used approach is to calculate affinities between a detection response $\boldsymbol{r}_i$ and a tracklet $\mathcal{T}_j$ in each cue (feature), and then to multiply all affinities to obtain the final association probability. In classic association frameworks [1]–[17], link probability between $\boldsymbol{r}_i$ and $\mathcal{T}_j$ is usually defined as the product of affinities based on several features, like position, size, appearance, etc., formulated as:

$$P_{link}(\boldsymbol{r}_i, \mathcal{T}_j) = A_{pos}(\boldsymbol{r}_i, \mathcal{T}_j) A_{sz}(\boldsymbol{r}_i, \mathcal{T}_j) A_{appr}(\boldsymbol{r}_i, \mathcal{T}_j) \cdots \quad (1)$$

where $A_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j)$ denote the affinity between a detection response $\boldsymbol{r}_i$ and a tracklet $\mathcal{T}_j$ and f_k denotes any feature used for target representation. However, as mentioned in Section I, multiplying affinities based on many features will not always increase discriminative power, on the contrary, it is prone to error accumulation and bring unnecessary computational cost. To address this problem, a novel hierarchical data association scheme is proposed.

B. Hierarchical Feature Space (HFS) Construction

First a feature space $\mathcal{F}$ is constructed which contains various common used features $\{f_k\}$ for describing detection responses. Based on the feature space $\mathcal{F}$, a generative form of the link probability is formulated as:

$$P_{link}(\boldsymbol{r}_i, \mathcal{T}_j | \mathcal{F}) = \prod_{f_k \in \mathcal{F}} \mathcal{A}_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j) \quad (2)$$

Then $\mathcal{F}$ is reconstructed into K hierarchies obeying two rules:

1) Lower hierarchies should be constructed with features which demonstrate higher reliability on target representation.
2) Higher hierarchy of feature space gradually have one more feature: $\mathcal{F}_{H_k} = \mathcal{F}_{H_{k-1}} \cup \{f_k\}$.

A brief illustration of hierarchical feature space is given in Fig. 2, where $\{\boldsymbol{r}^i_{H_k}\}$ and $\{T^j_{H_k}\}$ denote detection responses and tracklets to be associated in hierarchy H_k respectively.

C. Scene Adaptive Feature Selection (SAFS)

Therefore, to construct the hierarchical feature space, features with higher reliability should be selected first. It is based on the observation that reliability of a same feature varies a lot in different scenes. Reliable target representation is vital for an accurate data association. Therefore, the sence adaptive feature selection scheme is necessary, proposed as follows:

First, let $\boldsymbol{r}_i(t)$ denote detection response $\boldsymbol{r}_i$ at frame t, and tracklet $\mathcal{T}_j$ represents the set of all detection responses associated to target j before frame t. Suppose $\boldsymbol{r}_i$ is associated to $\mathcal{T}_j$ at frame t ($\boldsymbol{r}_i \rightarrow \boldsymbol{r}_j^{(t)}$), given feature f_k, the variation of feature representation for target j is defined as:

$$v^j_{f_k}(t) = Dist_{f_k}(\boldsymbol{r}_j^{(t)}, \mathcal{T}_j) \quad (3)$$

here $Dist_{f_k}(\cdot)$ represents distance metric between detection response $\boldsymbol{r}_i$ and target (tracklets) $\mathcal{T}_j$ under representation of feature f_k, which is user-defined metric but not fixed.

Then, two statistics of variation $v^i_{f_k}(t)$ are calculated: One is the mean of variation u_{f_k}:

$$u_{f_k} = \frac{\sum_{j=1}^{N_T} \sum_{l=1}^{t} \delta_j(l) v^j_{f_k}(l)}{\sum_{j=1}^{N_T} \sum_{l=1}^{t} \delta_j(l) + 1} \quad \text{where}$$
$$\delta_i(t) = \begin{cases} 1 & f_k \text{ is selected to represent } \boldsymbol{r}_{\mathrm{j}}^{(t)} \text{ on frame } t. \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

where N_T is number of detection responses associated to tracklets at time t. The other statistic of variation $v^i_{f_k}(t)$ is standard deviation of variation s_{f_k}:

$$s_{f_k} = \left(\frac{\sum_{j=1}^{N_T} \sum_{l=1}^{t} \left(\delta_j(l) v^j_{f_k}(l)\right)^2 - (u_{f_k})^2}{\sum_{i=1}^{N_T} \sum_{l=1}^{t} \delta_j(l) + 1} \right)^{\frac{1}{2}} \quad (5)$$

In practice, the ideal feature representation should have less and stable variation, that means low mean and standard deviation value. In this framework, the reliability of feature f_k on target representation can be given as:

$$R_k = U_k(\omega_1 \cdot \frac{1}{u_{f_k}} + \omega_2 \cdot \frac{1}{s_{f_k}}) \tag{6}$$

where U_k is a prior parameter for each feature f_k to transform the heterogeneous statistics of different features into homogeneous data to make sure reliability of different features can be compared together, and ω_1 and ω_2 are weights factors.

The formulations given in Eq. (4) and Eq. (5) are analytical forms. In practice, an iterative update way is adopted: given current u_{f_k}, s_{f_k} and new coming $v^i_{f_k}(t+1)$, the updated mean and standard deviation can be given as:

$$\hat{u}_{f_k} = \left(1 - \frac{1}{\Delta_n(t)+1}\right) \cdot u_{f_k} + \frac{1}{\Delta_n(t)+1} \cdot v^j_{f_k}(t+1) \tag{7}$$

$$\begin{aligned}(\hat{s}_{f_k})^2 &= \left(1 - \frac{1}{\Delta_n(t+1)+1}\right) \cdot (s_{f_k})^2 \\ &\quad + \frac{1}{\Delta_n(t+1)+1} \cdot \left(u^2_{f_k} - \hat{u}^2_{f_k} + \left(v^j_{f_k}(t+1)\right)^2\right) \\ &\text{where} \quad \Delta_n(t) = \sum_{j=1}^{N_T} \sum_{l=1}^{t} \delta_j(l)\end{aligned} \tag{8}$$

Given updated value $\hat{u}_{f_k}$ and $\hat{s}_{f_k}$, R_k can be iteratively updated by Eq. (6) at each frame t.

D. Hierarchical Data Association (HDA) on HFS

Based on previous modules, suppose K features with higher reliability are selected and hierarchical features space is constructed obeying rules in Section II-B. For kth hierarchy H_k, suppose there are M_k tracklets and N_k detection responses to be associated. Let $\mathcal{T}_{H_k} := \{\mathcal{T}^j_{H_k}\}_{M_k}$ denote M_k tracklets and $\mathcal{R}_{H_k} := \{\boldsymbol{r}^i_{H_k}\}_{N_k}$ denote N_k responses.

First, an affinity matrix $\mathcal{M}_{H_k}$ between $\mathcal{T}_{H_k}$ and $\mathcal{R}_{H_k}$ is calculated. Let $\mathcal{A}^{ij}_{H_k}$ denote the element in the ith row and jth column of $\mathcal{M}_{H_k}$. $\mathcal{A}^{ij}_{H_k}$ is the affinity between $\boldsymbol{r}_i$ and $\mathcal{T}_j$ considering all features $\{f_k\}$ in $\mathcal{F}_{H_k}$, given as:

$$\mathcal{A}^{ij}_{H_k} = P_{link}(\boldsymbol{r}_i, \mathcal{T}_j | \mathcal{F}_{H_k}) = \prod_{f_k \in \mathcal{F}_{H_k}} \mathcal{A}_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j) \tag{9}$$

$$\text{where} \quad \mathcal{A}_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j) = G(Dist_{f_k}(\boldsymbol{r}_i(t), \mathcal{T}_j); u_{f_k}, s^2_{f_k}) \tag{10}$$

Then, based on the affinity matrix $\mathcal{M}_{H_k}$, a hierarchical data association algorithm is conducted to handle data association on the kth hierarchy. A brief illustration of the hierarchical data association is shown in Fig. 2.

After that, reliable links in R_{H_k} are associated. For any noise detection in set N_{H_k}, a new tracklet is informally initialized first and will be formally initialized if enough responses are associated to it in subsequent frames. Thus new entry will be handled properly. For each tracklet $\mathcal{T}_j$ in miss detection set M_{H_k}, causes about miss are analyzed. If miss detection is due to exit, $\mathcal{T}_j$ is removed from $\mathcal{T}$. If due to occlusion, an occlusion handling strategy proposed in our previous work [18] is adopted. It can effectively find reappearing response and use it to update the tracklet $\mathcal{T}_j$. Conflicting links in set C_{H_k} are transferred to the higher hierarchy H_{k+1} to be further distinguished by combining more features in feature space $\mathcal{F}_{H_{k+1}}$. Finally, this iterative process is terminated until the last hierarchy H_K is processed or all conflicting links are distinguished. For example in Fig. 2, all conflicting links become reliable links in hierarchy H_3. The final remain conflicting links, if any, are associated using Nearest-Neighbour strategy for simplicity.

TABLE I
DETAILED CONFIGURATION OF EXPERIMENTAL DATASETS

	Dataset Type		
Dataset Name	In / Out	Distance	RGB-D / RGB
Scene 1 - Scene 5	Indoors	Close	RGB-D
Scene 6 - Scene 10	Indoors	Close	RGB
Scene 11 - Scene 15	Outdoors	Distant	RGB
PETS09 S2.L1 - S2.L3 [21]	Outdoors	Distant	RGB
TUD Campus [25]	Outdoors	Close	RGB
TUD Crossing [25]	Outdoors	Close	RGB

III. EXPERIMENTS AND ANALYSES

A. Datasets and Settings

There is no generally accepted benchmark available for multi-tracking [22]. Therefore, we recorded our own multi-tracking dataset including both RGB and RGB-D data, indoor and outdoor scenes, which shows great diversities. This challenging dataset presents frequent interactions, significant occlusions, various illumination conditions and cluttered backgrounds. The RGB-D dataset is recorded by a Kinect sensor. Height of the sensor is set to 1.8 m with a horizontal perspective. Moveover, most related publications have carried out experiments on their own sequences, which we have tried to combine several of them which are public. The detailed configuration of the experimental datasets is given in Table I.

B. Implementation Details

In practice, the reliability of features varies a lot in different scenes. To make the algorithm more general, scene adaptability is necessary. Compared with long period run of an online multi-tracking system, feature selection is a short period procedure, which does not need conducted in the whole process. For the initialization procedure in each scene, first a set of commonly used features are selected to construct the feature space, then data association is performed on one or more multi-tracking sequences recorded in this scene. Then based on the associated sequences, two statistics u_{f_k}, s_{f_k} and reliability R_k for all features in feature space are calculated according to algorithm given in Section II-C. Due to its statistical nature, occasional id-switches can be overlooked in data association. The variation reference U_k in Eq. (6) can be obtained by averaging variations of each feature f_k during association in several sequences in different scenes.

Our experiments are designed as follows: Firstly, quantitative analysis of features' reliability in different scenes are given which proves the reasonableness and necessity of the proposed scene adaptive feature selection scheme. Secondly, different multi-tracking frameworks with different feature selection schemes and different data association approaches are compared with quantitative and qualitative analysis, which proves the effectiveness and efficiency of the proposed framework.

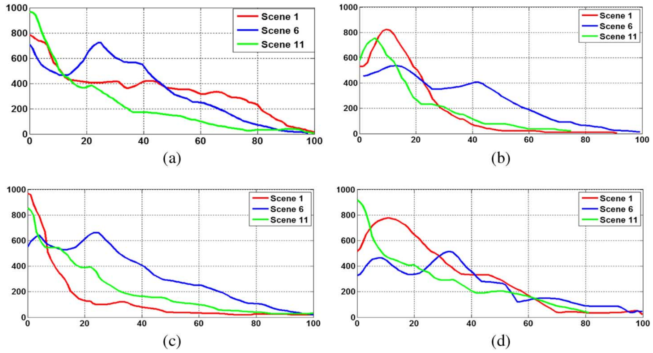

Fig. 3. Variation statistics of four features in different scenes. To facilitate observation, we mapped variation reference U_k (in Eq. (6)) of each feature f_k to 20 on horizontal axis, in order to give a intuitive observation of this comparison. The vertical axis indicates number of association corresponding to each variation value in the tested scenes.

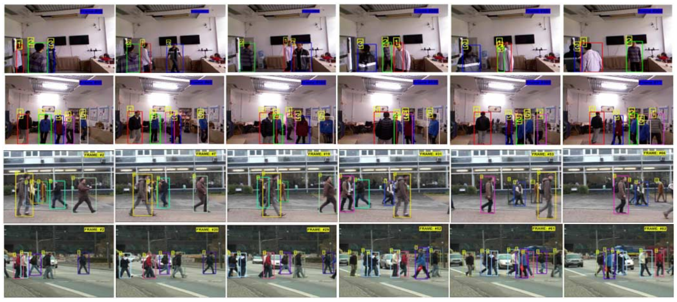

Fig. 4. Qualitative results of the proposed multi-tracking scheme tested in various scenes (from top to bottom scene 1, scene 6, TUD Campus [25] and TUD Crossing [25]).

C. Scene-Adaptive Feature Selection Evaluation

Three scenes in our datasets scene 1, scene 6 and scene 11 are used for the evaluation, which represents most typical scenes in practice, including indoor and outdoor, close and distant, sparse and crowded scenes, etc. (corresponds to row $1-3$ in Fig. 4 respectively). Four most commonly used features: color, position, size, and appearance models are selected for comparison. RGB and RGB-D datasets adopt different detector for getting detection responses for each frame. Here RGB dataset adopt classic DPM+LSVM detector [24] and RGB-D dataset adopt a novel 3D motion detection method [26]. Different detectors lead to different accuracy of detection responses. Therefore, common situations in practice such as various scenes, detectors and data types are included in our evaluation, based on common sense that these conditions relate with reliability of targets representation for a given feature.

To verify this, variation statistics of feature representation of four features in three different scenes are recorded during association and given in Fig. 3. In Fig. 3(a), crowded situation in scene 6 makes color representation (global HS histogram) less reliable compared with relatively sparse indoor and outdoor scenes in scene 1 and scene 11, for crowded scene brings more frequent occlusions which reduces stability of global HS histogram description. As shown in Fig. 3(c), size cue in scene 11 is much more reliable than in scene 6 for scale change and view-truncation are much severer in indoor scenes than outdoors. Size cue in scene 1 is most reliable for sizes of detection responses with RGB-D data are transformed to real length metric in real world which has little change for a same target.

TABLE II
QUANTITATIVE ACCURACY AND SPEED COMPARISON (NOT INCLUDING DETECTION TIME, ONLY DATA ASSOCIATION) OF SEVERAL DATA ASSOCIATION SCHEMES IN OUR DATASETS (TABLE I)

	MOTA (%) / FPS			
Frameworks	Scene 1-5	Scene 6-10	Scene 11-15	Public
SAFS+HDA	**91.20/29.3**	**86.23/20.5**	**89.20**/23.4	**88.25/21.9**
HDA	87.65/25.6	82.67/19.5	86.32/**23.6**	85.95/20.1
non-HDA	75.25/16.3	71.36/15.5	73.95/18.6	80.24/19.2

A similar situation with position cues is shown in Fig. 3(b) for both position and size are length-related features. For appearance, a commonly used part-based model in previous works [27]–[29] are adopted for scenes 6 and 11, but scene 1 uses a depth-invariant part-based model using RGB-D data which is proposed in our previous work [23] (details are given in the appendix Section III). Appearance representation is less reliable for scene 6 due to frequent view-truncation of the sensor, which seldom happens in outdoor scene 11. In conclusion, reliability of features closely relates to scene conditions (lighting, distance, density, etc.) and system conditions (accuracy of detectors, data source, such as RGB-D or RGB, etc.). This proves the rationality and necessity of the scene-adaptive scheme.

D. Hierarchical Data Association Evaluation

The standard metric MOTA (multiple objects tracking accuracy) [19] is widely used to evaluate performance of multi-tracking algorithm. In our experiments we adopted a revised version of MOTA following [20]. Taking into consideration of real-time capability of algorithm, we also compute the FPS. The quantitative results are given in Table II. Here non-HDA stands for combining all features in feature space for data association. It can be seen from Table II that compared with classic association schemes combining bunch of features, HDA not only improves the MOTA, but also the speed. This indicates that combining more features without considering requirements for current association not only wastes computation resources, but also weaken discriminant ability of discriminative representation of features. And with SAFS, the HDA is performed on the hierarchical feature spaces which gradually use more reliable features for association. It can be seen from Table II that both MOTA and speed are improved with SAFS. One more thing worth to mention is that speed is highly improved with RGB-D datasets because 3D position on the first hierarchy can solve most situations in multi-tracking.

IV. CONCLUSIONS

In this work, we focus on adaptively combining features to discriminate ambiguous targets for better data association in various scenes. Compared with previous work, the proposed hierarchical data association scheme based on hierarchical feature space gradually fuses more features according to requirements of distinguishing conflicting responses, leading to less error accumulation and less computational cost. The scene-adaptive scheme selects features with higher reliability in the applied scene based on the observation that features' reliability varies in different scenes and tracking systems. Experimental results demonstrates that scene-adaptive scheme is reasonable and necessary, and the proposed method contributes to an improvement in both accuracy and speed in multi-tracking. Future work will focus on learning more adaptive feature selection and more discriminative target representation for better data association.

References

[1] C. Kuo, C. Huang, and R. Nevatia, “Multi-target tracking by on-line learned discriminative appearance models,” in *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, 2010, pp. 685–692.

[2] Y. Li, C. Huang, and R. Nevatia, “Learning to associate: Hybridboosted multi-target tracker for crowded scene,” in *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, 2009, pp. 2953–2960.

[3] A. Andriyenko, K. Schindler, and S. Roth, “Discrete-continuous optimization for multi-target tracking,” in *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, 2012, pp. 1926–1933.

[4] J. Xing, H. Ai, L. Liu, and S. Lao, “Multiple player tracking in sports video: A dual-mode two-way bayesian inference approach with progressive observation modeling,” *IEEE Trans. Image Process.*, vol. 20, no. 6, pp. 1652–1667, 2011.

[5] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua, “Multiple object tracking using K-shortest paths optimization,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1806–1819, 2011.

[6] J. Berclaz, F. Fleuret, and P. Fua, “Robust people tracking with global trajectory optimization,” in *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, 2006, pp. 744–750.

[7] J. Xing, H. Ai, and S. Lao, “Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses,” in *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, 2009, pp. 1200–1207.

[8] L. Zhang, Y. Li, and R. Nevatia, “Global data association for multi-object tracking using network flows,” in *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, 2008, pp. 1–8.

[9] C. Huang, B. Wu, and R. Nevatia, “Robust object tracking by hierarchical association of detection responses,” in *Eur. Conf. Computer Vision, ECCV*, 2008, pp. 788–801.

[10] W. Choi, C. Pantofaru, and S. Savarese, “Detecting and tracking people using an RGB-D camera via multiple detector fusion,” in *IEEE Int. Conf. Computer Vision Workshops, ICCVW*, 2011, pp. 1076–1083.

[11] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool, “Robust tracking-by-detection using a detector confidence particle filter,” in *IEEE Int. Conf. Computer Vision, ICCV*, 2009, pp. 1515–1522.

[12] K. Okuma, A. Taleghani, O. D. Freitas, J. J. Little, and D. G. Lowe, “A boosted particle filter: Multitarget detection and tracking,” in *Eur. Conf. Computer Vision, ECCV*, 2004, pp. 28–39.

[13] B. Wu and R. Nevatia, “Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors,” *Int. J. Comput. Vis.*, vol. 75, no. 2, pp. 247–266, 2007.

[14] D. R. Magee, “Tracking multiple vehicles using foreground, background and motion models,” *Image Vis. Comput.*, vol. 22, no. 2, pp. 143–155, 2004.

[15] Z. Khan, T. Balch, and F. Dellaert, “MCMC-based particle filtering for tracking a variable number of interacting targets,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1805–1891, 2005.

[16] M. Yang, F. Lv, W. Xu, and Y. Gong, “Detection driven adaptive multi-cue integration for multiple human tracking,” in *IEEE Int. Conf. Computer Vision, ICCV*, 2009, pp. 1554–1561.

[17] B. Yang and R. Nevatia, “Online learned discriminative part-based appearance models for multi-human tracking,” in *Eur. Conf. Computer Vision, ECCV*, 2012, pp. 484–498.

[18] H. Liu, Y. Ze, and H. B. Zha, “Robust human tracking based on multi-cue integration and mean shift,” *Patt. Recognit. Lett.*, vol. 30, no. 9, pp. 827–837, 2009.

[19] K. Bernardin and R. Stiefelhagen, “Evaluating multiple object tracking performance: The clear mot metrics,” *EURASIP J. Image Video Process.*, vol. 1, pp. 1–10, 2008.

[20] H. B. Shitrit, J. Berclaz, F. Fleuret, and P. Fua, “Tracking multiple people under global appearance constraints,” in *IEEE Int. Conf. Computer Vision, ICCV*, 2011, pp. 137–144.

[21] J. Ferryman, in *IEEE Workshop Performance Evaluation of Tracking and Surveillance*, 2009.

[22] M. D. Breitenstein, F. Reichlin, B. Leibe, E. K. Meier, and L. V. Gool, “Online multiperson tracking-by-detection from a single, uncalibrated camera,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 33, no. 9, pp. 1820–1832, 2005.

[23] H. Liu and C. Wang, “Hierarchical data association and depth-invariant appearance model for indoor multiple objects tracking,” in *IEEE Int. Conf. Image Processing, ICIP*, 2013, pp. 2635–2639.

[24] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, “Object detection with discriminatively trained part based models,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, 2010.

[25] M. Andriluka, S. Roth, and B. Schiele, “People-tracking-by-detection and people-detection-by-tracking,” in *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, 2008, pp. 1515–1522.

[26] C. Wang, H. Liu, and L. Ma, “Depth motion detection—A novel RS-trigger temporal logic based method,” *IEEE Signal Process. Lett.*, 2014, 10.1109/LSP.2014.2313345 , to be published.

[27] C. X. Liu, S. G. Gong, C. C. Loy, and X. G. Lin, “Person reidentification: What features are important?,” in *Eur. Conf. Computer Vision Workshops, ECCVW*, 2012, pp. 391–401.

[28] B. Prosser, W. Zheng, S. Gong, and T. Xiang, “Person re-identification by support vector ranking,” in *Brit. Mach. Vis. Conf., BMVC*, 2010, vol. 1, no. 3, p. 5.

[29] W. S. Zheng, S. G. Gong, and T. Xiang, “Person re-identification by probabilistic relative distance comparison,” in *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*, 2011, pp. 649–656.