HIERARCHICAL DATA ASSOCIATION AND DEPTH-INVARIANT APPEARANCE MODEL FOR INDOOR MULTIPLE OBJECTS TRACKING

Hong Liu, Can Wang

Engineering Lab on Intelligent Perception for Internet of Things(ELIP),
Shenzhen Graduate School, Peking University, China

ABSTRACT

Discriminative target representation is vital for data association in multi-tracking. In order to increase the discriminative power, pervious works always combine bunch of features for target representation. However, this is prone to error accumulation and unnecessary computational cost, which may increase identity switches in data association on the contrary. To address this problem, we propose a hierarchical data association scheme which gradually combines features to the minimum requirements of discriminating ambiguous targets. In addition, indoor multi-tracking is more challenging due to frequent occlusion, view-truncation, large scale and pose variation, which may bring considerable unreliability for target representation. To handle this a novel depth-invariant part-based appearance model using RGB-D data is proposed . The depth-invariant appearance have stable length metric proportional to the absolute length metric in the world coordinates, which increase its robustness to scale variation. The part-based nature makes it robust to partial occlusion and view-truncation. Our algorithm is validated on various challenging indoor environments and it demonstrates high processing speed up to 50 fps and competitive accuracy.

Index Terms— Multiple Objects Tracking, Data Association, Appearance Model, RGB-D

1. 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. However, this is highly challenging in crowd environments with frequent occlusion, targets having similar appearances and complicated interaction. Most previous methods can be organized into two main categories: One category takes information from future frames [2, 3, 4, 5, 6] to get better association via global analysis, like global trajectory optimization[2], network flows [4], hierarchical tracklets association [7], 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 [9, 10, 11, 12]. They usually relied on Kalman [13] or particle filter [14] 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. In order to increase the system's discriminative power, many pervious works [1] [7] [15] combine a bunch of features to calculate affinities of detection responses in consecutive frames. But they are with unsatisfactory performance on handling relatively challenging indoor environments for two reasons. First, due to large appearance variation in indoor environments, detection responses of a same target may have large true variation. Examples of tough problems in practical indoor environments tracking is given in Fig.1. Second, cluttered environments inherently bring more observation errors for feature representation. Therefore combining a bunch of features to calculate affinities of detection responses is prone to severe error accumulation, and also bring unnecessary computational cost.

Fig. 1: Tough problems in indoor multi-tracking. The first row shows scenes with various illuminations. The bottom two rows are detection responses obtained from our dataset with large scale variation, frequent view-truncation, partial occlusion, and wider range of poses than outdoor pedestrians [8].

In order to address above problems and to achieve a time-critical indoor multi-tracking system, our work focus on efficiently combing features to discriminate ambiguous targets, and handling severe appearance variations in indoor environments. Our main contribution lies in two aspects: (1) A novel hierarchical scheme based on a self-constructed hierarchical feature space is proposed for data association. Features are gradually fused according to the need of discriminating ambiguous detection responses, this avoids unnecessary computation cost and reduce error accumulation compared to simultaneously fusing bunch of features; (2) A novel depth-invariant part-based appearance model is proposed to hand large scale variation and frequent view-truncation and partial occlusion in indoor environments. We validate our approach on a challenging dataset capturing various indoor scenes.

2. ASSOCIATION AND APPEARANCE MODEL

Motion Detection: Similar to many previous related works, data association is conducted on detection responses obtained by a detection method. In our framework, a simple but effective indoor mov-

This work is supported by National Natural Science Foundation of China(NSFC, No.60875050, 60675025), National High Technology Research and Development Program of China(863 Program, No.2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (No.JC201005280682A, CXC201104210010A).

ing objects detection method using RGB-D is adopted with real-time performance. But due to detection is not the focus of this article, it will not be elaborated here.

Preliminaries: Through out the paper, $\mathcal{R}^t := \{\boldsymbol{r}_i\}_n$ denotes n detection responses's set at frame t and $\boldsymbol{r}_i$ denotes one detection response. $\mathcal{T} := \{\mathcal{T}_j\}_m$ denotes m existing tracklets and $\mathcal{T}_j$ denotes one tracklet. In classic association frameworks [1][7][16], link probability between $\boldsymbol{r}_i$ and $\mathcal{T}_j$ is a kind of distance metric which is defined as the product of affinities based on several features, like position, size, appearance, etc., formulated as:

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

However, as mentioned in Section 1, 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:

2.1. Hierarchical Data Association on Hierarchical Feature Space

Hierarchical Feature Space: First a feature space $\mathcal{F}$ is constructed which contains various features $\{\boldsymbol{f}_k\}$ for describing detection responses. Based on the feature space $\mathcal{F}$, a generative form of the link probability considering observation errors $\boldsymbol{e}_{f_k}^{r_i}$,$\boldsymbol{e}_{f_k}^{T_j}$ and true variations $\boldsymbol{v}_{f_k}^{r_i,T_j}$ is formulated as:

$$P_{link}(\boldsymbol{r}_i, \mathcal{T}_j|\mathcal{F}) = \prod_{f_k \in \mathcal{F}} A_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j, \boldsymbol{e}_{f_k}^{r_i}, \boldsymbol{e}_{f_k}^{T_j}, \boldsymbol{v}_{f_k}^{r_i,T_j}) \quad (2)$$

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 in a tracking system. In other words, feature f_k should have smaller observation error $\boldsymbol{e}_{f_k}$, and smaller true variation $\boldsymbol{v}_{f_k}$.
2) The kth hierarchy of the feature space $\mathcal{F}_{H_k}$ contains all features in $\mathcal{F}_{H_{k-1}}$ and one more feature than $\mathcal{F}_{H_{k-1}}$: higher hierarchies gradually have more features.

Association on Hierarchical Feature Space: For kth hierarchy H_k, suppose there are M_k tracklets and N_k detection responses to be associated. Let $\mathcal{T}_{H_k} := \{\mathcal{T}_{H_k}^j\}_{M_k}$ be M_k tracklets' set and $\mathcal{R}_{H_k} := \{\boldsymbol{r}_{H_k}^i\}_{N_k}$ be N_k responses's set.

First, an affinity matrix $\mathcal{M}_{H_k}$ between $\mathcal{T}_{H_k}$ and $\mathcal{R}_{H_k}$ is calculated. Let $\mathcal{A}_{H_k}^{ij}$ denote the element in the ith row and jth column of $\mathcal{M}_{H_k}$. $\mathcal{A}_{H_k}^{ij}$ is the affinity between $\boldsymbol{r}_i$ and $\mathcal{T}_j$ considering all features $\{f_k\}$ in $\mathcal{F}_{H_k}$ (calculated via Eq.(2)).

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. The algorithm is given in the following pseudo-code table. Its function is to find reliable links R_{H_k}, conflicting links C_{H_k}, miss detections M_{H_k} and noise detections N_{H_k} in each hierarchy H_k.

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 [17] 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 $\mathcal{F}_{H_{k+1}}$.

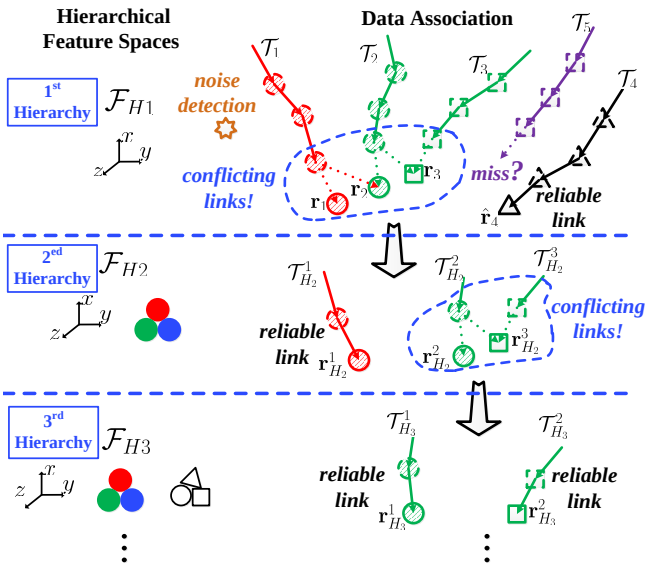

Fig. 2: A brief illustration of hierarchical data association on hierarchical feature space. Noise detection, miss detection, reliable links and conflicting links are unified and properly handled in this hierarchical framework.

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.

Feature Space Construction: In this framework, the feature space contains four features to describe detection responses. They are position $\boldsymbol{p}$, motion state $\boldsymbol{m}$, color $\boldsymbol{c}$ and appearance model $\boldsymbol{a}$. For each tracklet $\mathcal{T}_j$, its latest response $\boldsymbol{t}^{t-1}$ can be denoted as $\boldsymbol{t}^{t-1} = (\boldsymbol{p}_j^{t-1}, \boldsymbol{m}_j^{t-1}, \boldsymbol{c}_j, \boldsymbol{a}_j)$. Position $\boldsymbol{p}_j^{t-1}$ is represented as Euclidean location (X_c^j, Z_c^j) in camera coordinates in order to avoid perspective effect in 2D image plane, formulated as:

$$\boldsymbol{p}_j^{t-1} = (X_c^j, Z_c^j);\ Z_c^j = \alpha \cdot d_j;\ X_c^j = \frac{Z_c^j}{f_u}(u_j - u_0) \quad (3)$$

where d_j is depth of $\boldsymbol{t}_j^{t-1}$ and α is scale factor, f_u and u_0 is RGB camera's intrinsic parameters and u_j is RGB image's width coordinate. Position affinity between $\mathcal{T}_j$ and a new detection response $\boldsymbol{r}_i$ is calculated by:

$$\mathcal{A}_{\boldsymbol{p}}(\boldsymbol{r}_i|\mathcal{T}_j) = G(\|\boldsymbol{p}_i - \boldsymbol{p}_j^{t-1}\|; 0, \Sigma_{\boldsymbol{p}}) \quad (4)$$

Motion $\boldsymbol{m}_j$ of $\boldsymbol{t}_j^{t-1}$ in a tracklet $\mathcal{T}_j$ is formulated using First-order Markov Model as: $\boldsymbol{m}_j = \boldsymbol{p}_j^{t-1} - \boldsymbol{p}_j^{t-2}$. Motion affinity between $\mathcal{T}_j$ and $\boldsymbol{r}_i$ is:

$$\mathcal{A}_{\boldsymbol{m}}(\boldsymbol{r}_i|\mathcal{T}_j) = G(\|\boldsymbol{p}_i - \boldsymbol{p}_j^{t-1} - \boldsymbol{m}_j\|; 0, \Sigma_{\boldsymbol{m}}) \quad (5)$$

Color feature $\boldsymbol{c}_j$ is represented by the mean of 2D H-S histogram vectors of randomly selected responses belongs to $\mathcal{T}_j$ in HSV color

Algorithm 1 Hierarchical Data Association Algorithm.

Input: $\mathcal{M}_{H_k}, \mathcal{T}_{H_k}, \mathcal{R}_{H_k}$
Output: conflicting links set C_{H_k}; reliable links set R_{H_k}; miss detection set M_{H_k}; noisy detection set N_{H_k};

initial dual-threshold $\theta^1_{H_k}$ and $\theta^2_{H_k}$;
initial $C_{H_k}, M_{H_k}, R_{H_k}, N_{H_k}$ to $\varnothing$;
for $i = 1$ to N_k **do**
  $\mathcal{A}^{im^1_k}_{H_k} = \max\{\mathcal{M}^{i\cdot}_{H_k}\}, \mathcal{A}^{im^2_k}_{H_k} = \max\{\mathcal{M}^{i\cdot}_{H_k} - \{\mathcal{A}^{im^1_k}_{H_k}\}\}$;
  if $\mathcal{A}^{im^1_k}_{H_k} < \theta^1_{H_k}$ **then**
    $N_{H_k} = N_{H_k} + \{\boldsymbol{r}^i_{H_k}\}$;
  end if
  if $\mathcal{A}^{im^1_k}_{H_k} \geq \theta^1_{H_k}$ and $\mathcal{A}^{im^1_k}_{H_k} - \mathcal{A}^{im^2_k}_{H_k} \leq \theta^2_{H_k}$ **then**
    $C_{H_k} = C_{H_k} + \{\boldsymbol{r}^i_{H_k}\} + \{\mathcal{T}^{m^1_k}_{H_k}, \mathcal{T}^{m^2_k}_{H_k}\}$;
  end if
end for
for $j = 1$ to M_k **do**
  $\mathcal{A}^{n^1_k j}_{H_k} = \max\{\mathcal{M}^{\cdot j}_{H_k}\}, \mathcal{A}^{n^2_k j}_{H_k} = \max\{\mathcal{M}^{\cdot j}_{H_k} - \{\mathcal{A}^{n^1_k j}_{H_k}\}\}$;
  if $\mathcal{A}^{n^1_k j}_{H_k} < \theta^1_{H_k}$ **then**
    $M_{H_k} = M_{H_k} + \{\mathcal{T}^j_{H_k}\}$;
  end if
  if $\mathcal{A}^{n^1_k j}_{H_k} \geq \theta^1_{H_k}$ and $\mathcal{A}^{n^1_k j}_{H_k} - \mathcal{A}^{n^2_k j}_{H_k} \leq \theta^2_{H_k}$ **then**
    $C_{H_k} = C_{H_k} + \{\boldsymbol{r}^{n^1_k}_{H_k}, \boldsymbol{r}^{n^2_k}_{H_k}\} + \{\mathcal{T}^j_{H_k}\}$;
  end if
end for
$R_{H_k} = \{\mathcal{T}_{H_k} + \mathcal{R}_{H_k} - M_{H_k} - N_{H_k} - C_{H_k}\}$;

space for simplicity. Color affinity between $\mathcal{T}_j$ and $\boldsymbol{r}_i$ is:

$$\mathcal{A}_{\boldsymbol{c}}(\boldsymbol{r}_i|\mathcal{T}_j) = G\Big(\frac{1}{\text{corr}(\boldsymbol{c}_i, \boldsymbol{c}_j)}; 0, \Sigma_{\boldsymbol{c}}\Big) \tag{6}$$

where $corr(\boldsymbol{v}_i, \boldsymbol{v}_j)$ calculates correlation of vector $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$.

2.2. Depth-Invariant Part-Based Appearance Model

Depth Invariant Transform: In order to handle appearance variation due to partial occlusion and scale change caused by perspective effect, a novel depth-invariant part-based appearance model is proposed. Given any response $\boldsymbol{t}_j$ or $\boldsymbol{r}_j$, let $I_j = \{(u_i, v_i)\}$ denotes its RGB image patch inside the bounding box. Depth Invariant Transform (DIT) of I_j is formulated as:

$$\begin{aligned} u^i_{DI} &= \alpha' \cdot \big(W_{off} + d_i(u_i - u_0)\frac{1}{f_u}\big) \\ v^i_{DI} &= \alpha' \cdot \big(H_{off} + d_i(v_i - v_0)\frac{1}{f_v}\big) \end{aligned} \tag{7}$$

where (u_0, v_0, f_u, f_v) is RGB camera's intrinsic parameters and α' is scale factor. Then, I_j is transformed to a depth invariant patch $I^j_{DI} = \{(u^i_{DI}, v^i_{DI})\}$. Here, W_{off} is set to make sure the minimum value of u^i_{DI} is 1 and H_{off} is set to make sure the top of all depth invariant patches correspond to a same height in the world coordinates. For example, DIT procedure of two patches of detection responses of a same target are shown in Fig.3. The length metric in DIT patches is proportional to the absolute length metric in the world, and top of two patches (right) both correspond to 1.8m in the real world.

Part-based Model: The part-based model of $\boldsymbol{t}_j$ or $\boldsymbol{r}_j$ is further formulated as a concatenated vector $\boldsymbol{a}_j = (\boldsymbol{a}_{b1}, \boldsymbol{a}_{b2}, \cdots, \boldsymbol{a}_{bn})$, and each $\boldsymbol{a}_{bn}$ is the 2D H-S histogram vector of a part block and bn is number of part blocks. Different from pervious part-based approaches, as shown in right part of Fig.3, the block size of each part has a depth-invariant size (W^j_b, H^j_b), which correspond to fixed values in the world coordinates. This makes the appearance model tolerable to large scale changes. Moreover, in order to avoid interference of view-truncations, such as bottom-half body is often truncated by filed-of-view, or pose variations which is often caused by articulated arms and legs, only the top-half of a target is modeled. Appearance affinity between $\mathcal{T}_j$ and $\boldsymbol{r}_i$ is:

$$\mathcal{A}_{\boldsymbol{a}}(\boldsymbol{r}_i|\mathcal{T}_j) = G\Big(\frac{1}{\text{corr}(\boldsymbol{a}_i, \boldsymbol{a}_j)}; 0, \Sigma_{\boldsymbol{a}}\Big) \tag{8}$$

In this framework, features' affinities are all defined by Gaussian distributions and variances $\{\Sigma_{\boldsymbol{p}}, \Sigma_{\boldsymbol{m}}, \Sigma_{\boldsymbol{c}}, \Sigma_{\boldsymbol{a}}\}$ are given by experience according to the system's condition.

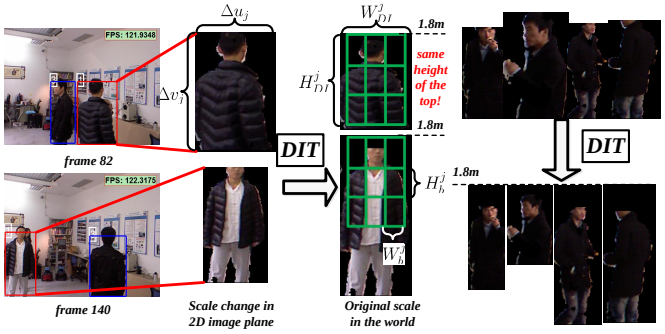

Fig. 3: Scale of the red target's detection responses changes in 2D image plane when target moves near and far, but it maintains relatively stable scale after Depth Invariant Transform (DIT). The right part shows patches of the blue target's detection responses before and after DIT during the tracking process.

3. EXPERIMENTS AND ANALYSES

Dataset: Our method is verified on a dataset of RGB-D video sequences capturing indoors multiple people walking from three scenes with room area from $16m^2$ to $36m^2$. This challenging dataset presents frequent interactions, significant occlusions, various illumination conditions and cluttered backgrounds (Fig.1 and Fig.4). The dataset is recorded by a Kinect sensor. Height of the sensor is set to $1.8m$ with a horizontal perspective. All experiments are conducted on an Intel $i5-2320$ 3.0 GHZ PC with 4.0 Gb RAM.

Implementation Details: For our approach, α and α' is set to 10 in Eq.(3) and Eq.(7). H_{off} is set to 100 in Eq.(7). 2D H-S histogram in color and appearance model are all set to 15 bins and 16 bins. For part-based appearance model, parts number bn is set to 9 and block size $\{H^j_b, W^j_b\}$ (shown in Fig.3) is set to $\{30, 20\}$, due to depth-invariant nature, these parameters correspond to absolute values in the world coordinates, they are relatively universal in the human tracking system. We empirically set $\{(\theta^k_1, \theta^k_2)\}$ in Algorithm 1 to $\{(0.8, 0.05), (0.6, 0.04), (0.45, 0.04), (0.3, 0.04)\}$ and set Gaussian distributions' variances $\{\Sigma_{\boldsymbol{p}}, \Sigma_{\boldsymbol{m}}, \Sigma_{\boldsymbol{c}}, \Sigma_{\boldsymbol{a}}\}$ to $\{50, 25, 0.2, 0.2\}$ according to system conditions.

Evaluation Metrics: The standard metric MOTA (multiple objects tracking accuracy) [18] is widely used to evaluate performance of multi-tracking algorithm which measures number of miss detections (MD), false positives (FP), and identity switches (IDS). In order to penalize more on identity switches, in our experiments we adopted a

revised version of MOTA following [19]. Taking into consideration of real-time capability of algorithm, we also compute the FPS.

Comparative evaluation: Three sets of comparison experiments are elaborated and analyses are given as follows:

First we evaluate advantages of our hierarchical feature space algorithm (HFSA) to the non-hierarchical all features algorithm (AFA). HFSA gradually fuses features layer-by-layer, while AFA fuses all features in one time. In the experiments, features used in AFA also have several combinations: one is position (P), motion (M), color (C), appearance model (A), used in our feature-space-construction, denoted as PMCA. Others are several combinations of position, size (S), motion, color, appearance widely used in previous related works. MOTA (%) and FPS of algorithms with several features combination (PMCA, PSMC, PSA) conducted on three scenes are given in Table 1. It shows AFA has relatively lower MOTA and lower processing speed. This is because combining all features for each association brings unnecessary computational cost and error accumulation. In addition, results in the $4-6$ rows shows features' combination also affects MOTA. This is because target representation based on a feature may have large true variations, such as size which is quite unreliable in indoor environments for large scale variation and view-truncation. Results in row 7 shows our system performs poorly with a different features order ACMP, for it violates the first rule of feature space construction: more reliable features should be fused first.

Table 1: Performance of hierarchical and non-hierarchical data association based on different combinations of features.

MOTA (%) / FPS	Scene 1	Scene 2	Scene 3
HFSA + PMCA	**88.5** /49.2	**92.0** /46.4	**89.5** /45.0
AFA + PMCA	87.5 /10.2	90.5 /11.6	83.2 /10.3
HFSA + PSMC	62.5 /–	65.4/–	57.2/–
AFA + PSMC	54.5 /–	56.4/–	50.5 /–
HFSA + PSA	64.5 /–	60.2/–	53.5/–
AFA + PSA	63.2 /–	59.0/–	52.3 /–
HFSA + ACPM	**62.5** /16.1	**60.0** /15.3	**57.5** /17.5

Second, we evaluate robustness of our depth-invariant part-based appearance model (DIPAM) in our dataset. Besides, comparative experiments are also conducted with classic average-division part-based appearance model (ADPAM) which divided responses bounding box into blocks averagely which is used in [16], and a state-of-the-art classifier-based discriminative appearance model (CDAM) proposed in [1] which combines color histograms, covariance matrixes, and histogram of gradients (HOG) for appearance modeling. DIPAM, ADPAM and CDAM are all fused in our hierarchical framework for data association. Experimental results are also given in form of MOTA (%) and FPS, shown in Table 2. ADPAM has low MOTA for it is vulnerable to large size variation compared with the absolute-length block size in DIPAM. The C-DAM performs well in MOTA for its classifier-based nature but is more computation-consuming in appearance modeling.

Finally, due to the recursive nature of our method, we evaluate the whole performance of our algorithm with two other classic recursive frameworks based on Kalman filter (KF) and partical filter (PF) in a 2030 frames sequence in our dataset, numbers of MD, FP and IDS is counted. We exactly implement two classic frameworks inspired by pervious works [13] and [14]. Quantitative results are given in Table 3. Compared with our method, KF represents higher identity switch rates for it cannot handle crowd multi-targets well and PF runs much slower for its nature of simultaneously exploring multiple hypotheses during tracking. Sampled tracking results in three scenes and tracking trails are given in Fig.4.

Table 2: Performance of our depth-invariant part-based appearance model and comparative appearance models.

MOTA (%) / FPS	Scene 1	Scene 2	Scene 3
DIPAM	**88.5** /49.2	**92.0** /46.4	**89.5** /45.0
ADPAM [16]	76.5 /50.2	80.5 /49.5	71.2 /42.0
CDAM [1]	87.5 /10.2	90.5 /11.6	83.2 /12.3

Table 3: Performance of our recursive framework and two other classic recursive frameworks.

Algorithms	MOTA (%)	MD	FP	IDS	FPS
Ours	**88.5**	**63**	**52**	**16**	**46.2**
KF [13]	54.2	245	150	98	15.2
PF [14]	69.5	105	98	45	3.2

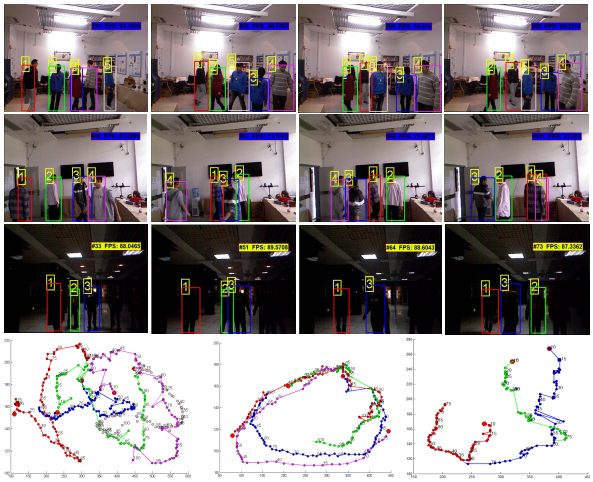

Fig. 4: Qualitative results of the proposed algorithm tested on our dataset. Bottom part shows three tracking trails of the three rows repectively.

4. CONCLUSIONS

In this work, we focus on efficiently combing features to discriminate ambiguous targets for better data association, and handling severe appearance variations in indoor environments. 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 lest computational cost. The novel depth-invariant part-based appearance model effectively handles large scale variation and frequent view-truncation and partial occlusion in indoor environments. As a result, our system demonstrates good performance in various challenging indoor scenes running in real-time. Future work will focus on learning more discriminative appearance models combining RGB-D data and apply it to mobile robot platforms.

5. REFERENCES

[1] C. Kuo, C. Huang, and R. Nevatia. "Multi-Target Tracking by On-Line Learned Discriminative Appearance Models". IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, pp. 685-692. 1, 2, 4

[2] J. Berclaz, F. Fleuret, and P. Fua. "Robust people tracking with global trajectory optimization". IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2006, pp. 744-750. 1

[3] J. Xing, H. Ai, and S. Lao. "Multi-object tracking through occlusions by local tracklets filtering and global tracklets association with detection responses". IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, pp.1200-1207. 1

[4] L. Zhang, Y. Li, and R. Nevatia. "Global data association for multi-object tracking using network flows". IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1-8. 1

[5] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. "Tracking Multiple People under Global Appearance Constraints". IEEE International Conference on Computer Vision, ICCV 2011, pp. 137-144. 1

[6] J. Berclaz, F. Fleuret, E. Turetken, and P. Fua. "Multiple Object Tracking using K-Shortest Paths Optimization". IEEE Transaction on Pattern Analysis and Machine Intelligence, TPAMI Sept. 2011, Vol. 33(9), pp. 1806-1819. 1

[7] C. Huang, B. Wu, and R. Nevatia. "Robust Object Tracking by Hierarchical Association of Detection Responses". European Conference on Computer Vision, ECCV 2008, pp. 788-801. 1, 2

[8] W. Choi, C. Pantofaru and S. Savarese. "Detecting and Tracking People using an RGB-D Camera via Multiple Detector Fusion". IEEE Internation Conference on Computer Vision Workshops, ICCVW 2011. 1

[9] M. D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, and L. V. Gool. "Robust tracking-by-detection using a detector confidence particle filter". IEEE International Conference on Computer Vision, ICCV 2009, pp. 1515-1522. 1

[10] Y. Cai, N. de Freitas, and J. J. Little. "Robust visual tracking for multiple targets". European Conference on Computer Vision, ECCV 2006, pp. 107-118. 1

[11] K. Okuma, A. Taleghani, O. D. Freitas, J. J. Little, and D. G. Lowe. "A boosted particle filter: Multitarget detection and tracking". European Conference on Computer Vision, ECCV 2004, pp. 28-39. 1

[12] B. Wu, and R. Nevatia. "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors". International Journal of Computer Vision, IJCV Nov. 2007, Vol. 75(2), pp. 247-266. 1

[13] D. R. Magee. "Tracking Multiple Vehicles Using Foreground, Background and Motion Models". Image and Vision Computing, 2004, pp. 143-155. 1, 4

[14] Z. Khan, T. Balch, and F. Dellaert. "MCMC-Based Particle Filtering for Tracking a Variable Number of Interacting Targets". IEEE Transaction on Pattern Analysis and Machine Intelligence, TPAMI Nov. 2005, Vol. 27(11), pp. 1805-1891. 1, 4

[15] M. Yang, F. Lv, W. Xu, and Y. Gong. "Detection driven adaptive multi-cue integration for multiple human tracking". IEEE International Conference on Computer Vision, ICCV 2009, pp. 1554-1561. 1

[16] B. Yang, and R. Nevatia. "Online Learned Discriminative Part-Based Appearance Models for Multi-Human Tracking". European Conference on Computer Vision, ECCV 2012. 2, 4

[17] H. Liu, Y. Ze, H. B. Zha. "Robust Human Tracking Based on Multi-Cue Integration and Mean Shift". Pattern Recognition Letters, PRL July 2009, Vol. 30(9), pp. 827-837. 2

[18] K. Bernardin and R. Stiefelhagen. "Evaluating Multiple Object Tracking Performance: the Clear Mot Metrics". EURASIP Journal on Image and Video Processing, JIVP Jan. 2008, Vol. 2008(1), pp. 1-10. 3

[19] H.B. Shitrit, J. Berclaz, F. Fleuret and P. Fua. "Tracking multiple people under global appearance constraints". EEE International Conference on Computer Vision, ICCV 2011, pp. 137-144. 4