Available online at www.sciencedirect.com

ScienceDirect

CAAI Transactions on Intelligence Technology 1 (2016) 210–224
http://www.journals.elsevier.com/caai-transactions-on-intelligence-technology/

Original Article

Scene-adaptive hierarchical data association and depth-invariant part-based appearance model for indoor multiple objects tracking☆

Hong Liu*, Can Wang, Yuan Gao

Key Laboratory of Machine Perception (Ministry of Education) and Engineering Lab on Intelligent Perception for Internet of Things (ELIP), Shenzhen Graduate School, Peking University, China

Available online 1 November 2016

Abstract

Indoor multi-tracking is more challenging compared with outdoor tasks due to frequent occlusion, view-truncation, severe scale change and pose variation, which may bring considerable unreliability and ambiguity to target representation and data association. So discriminative and reliable target representation is vital for accurate data association in multi-tracking. Pervious works always combine bunch of features to increase the discriminative power, but this is prone to error accumulation and unnecessary computational cost, which may increase ambiguity on the contrary. Moreover, reliability of a same feature in different scenes may vary a lot, especially for currently widespread network cameras, which are settled in various and complex indoor scenes, previous fixed feature selection schemes cannot meet general requirements. To properly handle these problems, first, we propose a scene-adaptive hierarchical data association scheme, which adaptively selects features with higher reliability on target representation in the applied scene, and gradually combines features to the minimum requirement of discriminating ambiguous targets; second, a novel depth-invariant part-based appearance model using RGB-D data is proposed which makes the appearance model robust to scale change, partial occlusion and view-truncation. The introduce of RGB-D data increases the diversity of features, which provides more types of features for feature selection in data association and enhances the final multi-tracking performance. We validate our method from several aspects including scene-adaptive feature selection scheme, hierarchical data association scheme and RGB-D based appearance modeling scheme in various indoor scenes, which demonstrates its effectiveness and efficiency on improving multi-tracking performances in various indoor scenes.

Keywords: Multiple objects tracking; Scene-adaptive; Data association; Appearance model; RGB-D data

1. Introduction

In recent years, due to the growing demand for smart home applications [1–3], issues like multi-tracking [4,5], and re-identification [6–8] attract more and more attention from researchers in computer vision field, and also have many problems to be solved. Multiple objects tracking has been an active research topic in computer vision within a long period of time. It aims to locate moving objects, maintain their identities and retrieve their trajectories [4]. However, this is highly challenging in crowd environments with frequent occlusion, targets having similar appearances and complicated interaction.

Most previous methods that focus on multiple objects tracking can be organized into two main categories: One category takes information from future frames [9–13] to get better association via global analysis, like global trajectory optimization [9], network flows [11], hierarchical tracklets association [14], etc. However, they are based on the prerequisite that detection responses in all frames are given, both

☆ This work is supported by National Natural Science Foundation of China (NSFC, No. 61340046), National High Technology Research and Development Program of China (863 Program, No. 2006AA04Z247), Scientific and Technical Innovation Commission of Shenzhen Municipality (JCYJ20130331144631730, JCYJ20130331144716089), Specialized Research Fund for the Doctoral Program of Higher Education (No. 20130001110011).
* Corresponding author.
E-mail addresses: hongliu@pku.edu.cn (H. Liu), canwang@pku.edu.cn (C. Wang), ygao@sz.pku.edu.cn (Y. Gao).
Peer review under responsibility of Chongqing University of Technology.

http://dx.doi.org/10.1016/j.trit.2016.10.008

from past and future, so it is not suitable for time-critical applications and is relatively computation-consuming when performing a global optimization. The other category only considers past and current frames to make association decisions [15–18]. They usually relies on Kalman [19] or particle filter [20] to handle data association. Because of their recursive nature, this category is suitable for time-critical applications, 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 current frame, but also has stable representation for each target in consecutive frames.

In order to increase the discriminating power, many pervious works [4,14,22] usually combine a bunch of features to represent detection responses and calculate the affinity matrix between them and existing tracklets. But they are with unsatisfactory performances 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 the target is hard to obtain. Second, observation errors of targets representation are common in cluttered scenes. For example, positions of detection responses may not be exactly on the center of targets due to frequent view-truncation and partial occlusion (as shown in Fig. 1), especially in indoor scenes field-of-view of sensors is relatively limited compared with outdoor scenes. In addition, accuracy of a detection response also relies on the detector's performance in the applied scene, which may vary a lot in different scenes. This leads to that same feature representation may have different reliability in different scenes. Therefore, combining a bunch of features may not contribute to a better association between detection responses and existing tracklets. On the contrary, features have lower reliability or discriminating power may bring adverse effect on reliable and discriminating features, and also unnecessary computational cost. Moreover, currently applications on network cameras harshly require more general and scene-adaptive schemes to handle the variety and diversity of the widespread scenes.

Therefore, motivated by properly handling the above problems and in order to achieve a time-critical indoor multi-tracking system, our work focus on accurate data association, scene-adaptive feature selection, and better appearance model. Our main contribution lies in three aspects: First, a novel hierarchical data association scheme based on the hierarchical feature space is proposed. Features are gradually combined during data association procedure 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; Second, a scene-adaptive feature selection scheme is proposed, which measures features' reliability and disciminability of features used for targets representation in the applied scene, and selects relatively reliable and discriminative features for data association. This makes the algorithm much more general for various scenes in the camera network; Third, a novel depth-invariant appearance model is proposed as a high level feature for target representation, which properly handles server scale problem on 2D image plane, frequent view-truncation and partial occlusion which occur commonly in indoor environments. Experiments conducted in a variety of scenes demonstrate the effectiveness of reliable feature selection and hierarchical data association. The depth-invariant appearance model based on RGB-D data also shows its effectiveness when dealing with occlusion and scale change in complex indoor scenes.

2. Related work

2.1. Data association

Data association based multi-tracking methods become increasingly popular driven by the recent progress in object

Fig. 1. Tough problems for multi-tracking tasks. The first row shows various indoor and outdoor scenes with various illumination conditions, crowdedness and angle of views. The bottom two rows are detection responses obtained from our indoor datasets with large scale variation, frequent truncation by the field-of-view, partial occlusion, and wider range of poses than outdoor pedestrians [21].

detection and pedestrian detection [23], which links current detection responses to existing trajectories during multi-tracking. Recently, more and more works adopt data association based tracking scheme to implement robust multi-tracking [24–26] while utilizing different frameworks. The data association based tracking scheme treats every detection response as a data unit and try to association these data units between consecutive frames which belong to the same target. Based on the detection responses in each frame, most related works utilizes future frames to get better association via global analysis, like global trajectory optimization [9] and network flows [11]. This increases the accuracy of the data association due to introducing of future information, but it is not suitable for a time-critical application. Therefore, in this work in order to achieve a time-critical multi-tracking system, we tend to only use past and current detection responses to perform data association online, which has a requirement for more accurate data association in each current frame. Obviously, feature representation of detection responses is vital for the accuracy of data association, so a reliable feature selection scheme is important and is one of the main focus in this work.

2.2. Feature representation and appearance models

Feature representation of targets is crucial in data association because it determines whether target representation is able to distinguish one target from others and maintains its own identification along the time sequence. Previous researchers tend to use basic features to achieve a generative description for each target [14,27], like color, size, motion, position, etc., and our framework also follows this scheme for better practical performance. Different from these classic features, appearance model can be regarded as a high level feature which is vital for an accurate representation of detection response. Many researchers are devoted to more sophisticated appearance models to represent detection responses [4,28,29] in the field of multi-tracking.

Given these basic features, how to combine them with spatial information determines the effectiveness of an appearance model. Undoubtedly, spatial information plays a crucial role on describing appearances because it essentially implies the spatial layout of basic features. Work in [31] equally subdivides the bounding box of detection response into ten horizontal stripes, and work in [28] uses fifteen squares to build part-based appearance models so that the parts are not too large to model occlusions and not too small to include meaningful features. In addition to these simple and intuitive spatial layouts, work in [30] further combines spatial symmetry of human body to hand view variation. Although previous works try various spatial layout schemes for better introducing spatial information, they have essential defects because the spatial information they used is purely based on the image patch on 2D image plane. But in practical situations detected regions always exhibit dislocation, truncation or occlusion, caused by unsatisfied detector performance in complicated environments. Therefore, in order to handle this problem, we propose a novel depth-invariant part-based appearance model based on RGB-D data, which will be described in details in Section 4.

2.3. RGB-D data for indoors applications

Recently, combining RGB and depth data for computer vision applications becomes more and more popular [37], because recent emergence of depth sensor (e.g., Microsoft Kinect) has made it feasible and economically sound to capture in real-time not only color images but also depth maps with appropriate resolution and accuracy [33]. The introduce of depth data provides more information from the third dimension which reduces troubles which are caused by the ill-posed problem in 2D image processing. Moreover, the depth data is seldom affected by bad illumination conditions which are common in indoors environments, which can be a complementary cue of RGB data.

Therefore, in order to handle more practical challenges and achieve a time-critical indoor multi-tracking system, depth data is adopted in our work. Previous works adopt depth data for motion detection [32], background subtraction [38] and 3D body pose estimation [34], which provides robust and various target detection schemes. Many depth-based features are also proposed by previous works [35,36] for both target detection and representation. In this work, based on depth data various depth-based cues are used for target representation, such as 3D position, 3D motion and 3D spatial layouts of basic features for appearance modeling. The introduction of these depth-based features increases diversity of features, which provides more reliable features for target representation. This highly improves robustness of the system in practical application with problems of cluttered environment, bad illumination conditions and scale variation on 2D image plane.

3. Scene-adaptive hierarchical data association

3.1. 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 $\mathscr{R}^t := \{\boldsymbol{r}_i\}_n$ 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, formulated as:

$$\mathcal{T}_j := \left\{\cdots, \boldsymbol{r}_j^{(t-2)}, \boldsymbol{r}_j^{(t-1)}\right\} \tag{1}$$

here $\boldsymbol{r}_j^{(t-1)}$ denotes the detection response associated to tracklet $\mathcal{T}_j$ at time $t-1$. It can be seen that one tracklet actually contains all detection responses associated to it in the past. At time t during data association, for each new detection response $\boldsymbol{r}_i^{(t)}$, we should associate it to its corresponding tracklet $\mathcal{T}_j$.

The most commonly used approach is to calculate affinities between a detection response $\boldsymbol{r}_i$ and a tracklet $\mathcal{T}_j$ under representation of one or more features, and then to multiply all affinities to obtain the final association probability.

In classic association frameworks [4,14,28], 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:

$$P_{link}(\boldsymbol{r}_i, \mathcal{T}_j) = \mathcal{A}_{pos}(\boldsymbol{r}_i, \mathcal{T}_j)\mathcal{A}_{sz}(\boldsymbol{r}_i, \mathcal{T}_j)\mathcal{A}_{ap}(\boldsymbol{r}_i, \mathcal{T}_j)\cdots \quad (2)$$

where $\mathcal{A}(\boldsymbol{r}_i, \mathcal{T}_j)$ denotes the affinity between a detection response $\boldsymbol{r}_i$ and a tracklet $\mathcal{T}_j$, and subscribes *'pos'*, *'sz'* and *'ap'* denotes position, size and appearance features used for target representation respectively. For each association, previous work always combines these features to calculate the link probability in order to increase responses' discriminability. It seems make sense and do achieve good results in several literatures. However, practical experiments and analysis show that multiplying affinities based on many features will not always increase discriminative power, on the contrary, it is prone to error accumulation from multiple feature representations and brings unnecessary computational cost.

A brief example of how observation errors affect data association is given in Fig. 2. Let $\overline{\mathcal{T}}_1$ and $\overline{\mathcal{T}}_2$ denote true values of two tracklets in a given feature space, in other words under representation of a given feature f_k. Because any feature representation can be regarded as a data point in its own feature space. Let $\mathcal{T}_1$ and $\mathcal{T}_2$ denote their observed values. Here $\boldsymbol{r}_i$ and $\boldsymbol{r}_l$ denote two detection responses detected in the current frame, and $\bar{\boldsymbol{r}}_i$ and $\bar{\boldsymbol{r}}_l$ are their corresponding true values. It can be seen that the observed affinities between $\mathcal{T}_1$ and two responses $\boldsymbol{r}_i$ and $\boldsymbol{r}_l$ are almost the same, but the true values differ a lot. This is due to observation errors $e^{r_i}_{f_k}$, $e^{r_l}_{f_k}$ and $e^{\mathcal{T}_1}_{f_k}$ between observed detection responses and their true positions in the given feature space.

Unfortunately, this is the case for almost all feature representations in practical application. This indicates that if even non-trivial observation errors exist under several feature representation, they will accumulate when multiplying feature affinities, as the classic works did in Formula (2). Therefore, the final link probability P_{link} in Formula (2) may not reflect the real affinity between detection responses and tracklets. Even there exists reliable and discriminating feature representations with less observation error, their discriminating power can be considerably reduced due to observation error accumulation. On the other side, if true variation under a certain feature representation is large, even the detection response and tracklet belong to a same target, the affinity may be small. Therefore, this requires a more reasonable data association scheme with reliable feature representation, which is the main focus in this work.

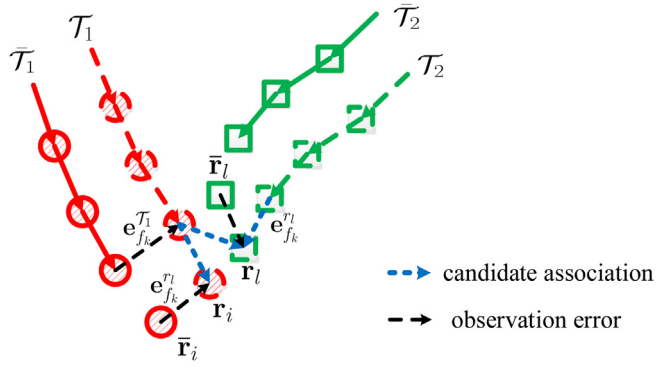


Fig. 2. A simple example of how observation errors affect data association.

3.2. Hierarchical data association

As mentioned above, combining a bunch of features may not contribute to a better association between detection responses and existing tracklets. On the contrary, features have lower reliability or discriminating power can bring adverse effect on other reliable and discriminating features, and also unnecessary computational cost. These problems are relatively severe in indoor multiple objects tracking for two reasons: first, the most commonly used features in classic methods such as position, size, color and appearance model are easily affected by partial occlusion, view-truncation and bad illumination, which are common situations in indoor scenes, so this may bring larger observation errors to targets representation; second, under these complicated situations, even detection responses belong to a same target may exist larger variation. In order to handle these problems, a novel hierarchical data association scheme is proposed as follows:

3.2.1. Hierarchical feature space construction

First, a feature space $\mathcal{F}$ is constructed with 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:

$$P_{link}(\boldsymbol{r}_i, \mathcal{T}_j|\mathcal{F}) = \prod_{f_k \in \mathcal{F}} \mathcal{A}_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j) \quad (3)$$

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

1) Lower hierarchies of the feature space should be constructed with features which demonstrate higher reliability on target representation. In other words, target representation based on feature f_k should have smaller observation errors and true variations.
2) Higher hierarchies of the feature space gradually have one more feature compared with the lower ones, which can be formulated as $\mathcal{F}_{H_k} = \mathcal{F}_{H_{k-1}} \cup \{f_k\}$.

A brief illustration of hierarchical feature space is given in Fig. 3, where $\{\boldsymbol{r}^i_{H_k}\}$ and $\{T^j_{H_k}\}$ denote detection responses and tracklets to be associated in hierarchy H_k respectively. Based on the hierarchical data association scheme, 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.

3.2.2. Data association on HFS

Based on previous modules, suppose K features with higher reliablity are selected and hierarchical features space is

constructed obeying rules in Section 3.2. 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:

$$\mathcal{A}^{ij}_{H_k} = P_{link}(\boldsymbol{r}_i, \mathcal{T}_j | \mathcal{F}_{H_k}) = \prod_{f_k \in \mathcal{F}_{H_k}} \mathcal{A}_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j) \tag{4}$$

$$\text{where} \quad \mathcal{A}_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j) = G(\mathcal{D}_{f_k}(\boldsymbol{r}_i(t), \mathcal{T}_j); \mu_{f_k}, \Sigma_{f_k}) \tag{5}$$

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 pseudo-code procedure in Algorithm 1. 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. A brief illustration of the hierarchical data association is shown in Fig. 3.

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 [47] 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. 3, all conflicting links become reliable links in hierarchy H_3. The final remaining conflicting links, if any, are associated using Nearest-Neighbor strategy for simplicity.

Therefore, through this hierarchical data association scheme, on one hand features with higher reliability can be firstly used for data association, which reduces observation error accumulation compared with the classic schemes. On the other hand several practical multi-tracking problems such as miss detection, noise detection can be handled in this unified framework.

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**
    $C_{H_k} = C_{H_k} + \{\boldsymbol{r}^{n^1_k}_{H_k}, \boldsymbol{r}^{n^2_k}_{H_k}\} + \{\mathcal{T}^j_{H_k}\}$;
  else
    $R_{H_k} = R_{H_k} + \{\boldsymbol{r}^{n^1_k}_{H_k}\} + \{\mathcal{T}^j_{H_k}\}$;
  end if
end for
$R_{H_k} = \{\mathcal{T}_{H_k} + \mathcal{R}_{H_k} - M_{H_k} - N_{H_k} - C_{H_k}\}$;

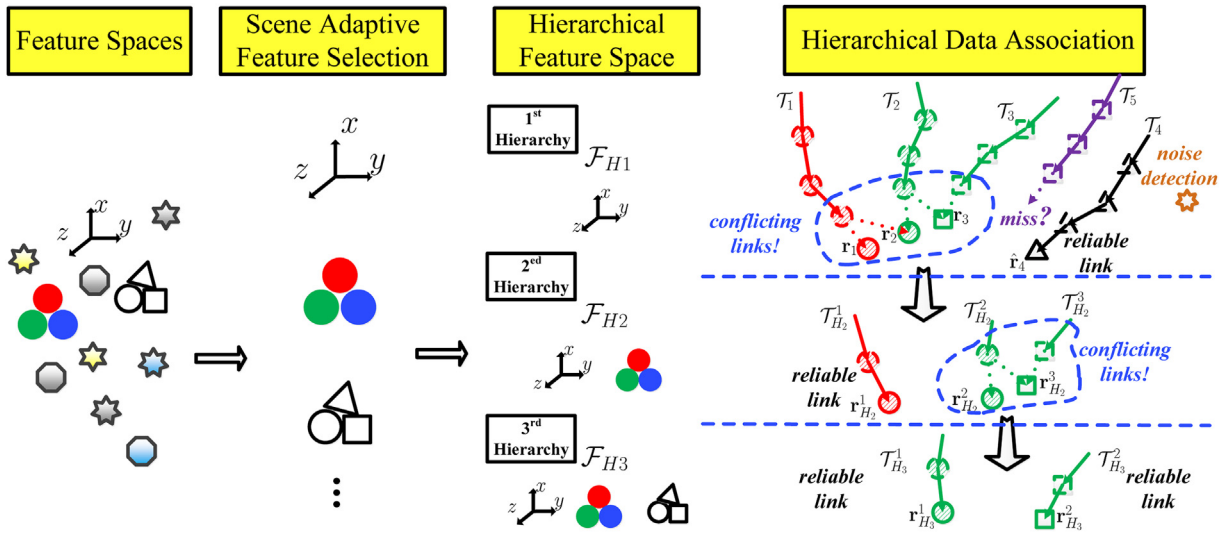


Fig. 3. A brief illustration of hierarchical data association on scene-adaptive feature space. Noise detection, miss detection, reliable links and conflicting links are unified and properly handled in this hierarchical framework.

3.2.3. Feature representation

In this part, details of feature space construction will be elaborated. The classic features widely used in multi-tracking such as position, size, color, appearance model [14,27], are all adopted in this work, because they are essentially suitable for the nature of multi-tracking problem. The difference of this work is that with RGB-D data, the feature representation design is different from RGB data, which combines more spatial information compared with the RGB version. They can get more accurate and reliable description when combining depth data. Here we focus on introducing the RGB-D version of classic features we used for feature space construction, such as position $\boldsymbol{p}$, motion state $\boldsymbol{m}$, color $\boldsymbol{c}$, appearance model $\boldsymbol{a}$ etc., and their general formulations are elaborated as follows:

Suppose for each tracklet $\mathcal{T}_j$, its latest response $\boldsymbol{r}_j^{t-1}$ can be denoted as:

$$\boldsymbol{r}_j^{t-1} = \left\{\boldsymbol{p}_j^{t-1}, \boldsymbol{m}_j^{t-1}, \boldsymbol{c}_j, \boldsymbol{a}_j, \cdots\right\} \tag{6}$$

where $\boldsymbol{p}_j^{t-1}$, $\boldsymbol{m}_j^{t-1}$, $\boldsymbol{c}_j$ and $\boldsymbol{a}_j$ indicate features of position, motion, color and appearance model respectively.

Position is almost the most important and is widely used feature in multi-tracking. Besides the 2D version position, which is always set as the center or top of the detection response on 2D image plane, we utilize depth information and adopt its 3D version. Position $\boldsymbol{p}_j^{t-1}$ is represented as Euclidean location (X_c^j, Z_c^j) in camera coordinates based on depth data, in order to avoid perspective effect in 2D image plane, formulated as:

$$\boldsymbol{p}_j^{t-1} = \left(X_c^j, Z_c^j\right);\ Z_c^j = \alpha_d \cdot d_j;\ X_c^j = \frac{Z_c^j}{f_u}\left(u_j - u_0\right) \tag{7}$$

where d_j is average depth of the detection response $\boldsymbol{r}_j^{t-1}$ on depth image and α_d is scale factor for quantization. Here f_u and u_0 is the intrinsic parameters of the camera and u_j is the width coordinate on 2D image plane.

As mentioned in the last section, given any detection response $\boldsymbol{r}_i$ and tracklet $\mathcal{T}_j$, a user-defined metric $F_{f_k}(\cdot)$ is used for the calculation of affinity $\mathcal{A}_{f_k}(\boldsymbol{r}_i, \mathcal{T}_j)$ based on each feature f_k. Here, similar to previous related works such as [14], position affinity is assumed to obey a Gaussian distribution, which is also validated by observation in practice. So position affinity between $\mathcal{T}_j$ and a new detection response $\boldsymbol{r}_i$ is calculated by:

$$\mathcal{A}_{\boldsymbol{p}}\left(\boldsymbol{r}_i, \mathcal{T}_j\right) = G\left(\left\|\boldsymbol{p}_i - \boldsymbol{p}_j^{t-1}\right\|; \mu_{\boldsymbol{p}}, \Sigma_{\boldsymbol{p}}\right) \tag{8}$$

where $\mu_{\boldsymbol{p}}$ and $\Sigma_{\boldsymbol{p}}$ are scene-adaptive parameters according to reliability of features in each scene, which will elaborated in detail in Section 3.3.

Similarly, details of motion cue and its affinity distance metric is given as follows. 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 also given obeying Gaussian distribution with its own scene-adaptive parameters $\mu_{\boldsymbol{m}}$ and $\Sigma_{\boldsymbol{m}}$, formulated as:

$$\mathcal{A}_{\boldsymbol{m}}\left(\boldsymbol{r}_i, \mathcal{T}_j\right) = G\left(\left\|\boldsymbol{p}_i - \boldsymbol{p}_j^{t-1} - \boldsymbol{m}_j\right\|; \mu_{\boldsymbol{m}}, \Sigma_{\boldsymbol{m}}\right) \tag{9}$$

Different from location-based features such as position and motion, content-based features such color histogram and appearance models are wildly used for targets representation in multi-tracking.

Color feature $\boldsymbol{c}_j$ can be designed as histogram vectors of some commonly used color space such as HSV, RGB, YUV etc., or combinations of them. Similarly, the metric of color affinity between $\mathcal{T}_j$ and $\boldsymbol{r}_i$ is given by follows:

$$\mathcal{A}_{\boldsymbol{c}}\left(\boldsymbol{r}_i, \mathcal{T}_j\right) = G\left(\frac{1}{\mathrm{corr}\left(\boldsymbol{c}_i, \boldsymbol{c}_j\right)}; \mu_{\boldsymbol{c}}, \Sigma_{\boldsymbol{c}}\right) \tag{10}$$

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

For appearance model, it is also regarded as a kind of high level features in the feature space. Due to its sophisticated design, it will be elaborated in details in Section 4.

Besides these features mentioned above, other widely used features such as textures, edges, Haar-like features, interest points, image patches, segmented regions [44] and also the combination of these basic features [45] [46] are also included in our feature pool.

3.3. Scene adaptive feature selection (SAFS)

As mentioned before, reliability of a same feature in different scenes may vary a lot, especially for currently widespread network cameras, so previous fixed feature selection schemes cannot meet general requirements. In this section, we propose a scene adaptive feature selection scheme which helps to select more reliable feature in each scene and use them to construct the hierarchical feature space.

Obeying two rules proposed in Section 3.2.1, 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, but a reliable target representation is vital for an accurate data association. Therefore, the scene adaptive feature selection scheme is necessary, proposed as follows:

First, for the convenience of the following description, 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, written as $\mathcal{T}_j := \{\cdots, \boldsymbol{r}_j^{(t-2)}, \boldsymbol{r}_j^{(t-1)}\}$. So if $\boldsymbol{r}_i$ is associated to $\mathcal{T}_j$ at frame t, then $\boldsymbol{r}_i$ is also can be written as $\boldsymbol{r}_j^{(t)}$.

If $\boldsymbol{r}_i$ has already been associated to $\mathcal{T}_j$ at frame t, given feature f_k, we are interested in whether the feature representation is stable between $\boldsymbol{r}_j^{(t)}$ and $\boldsymbol{r}_j^{(t-1)}$, which is vital for data association. The variation between them is defined as:

$$v_{f_k}^j(t) = \mathcal{D}_{f_k}\left(r_j^{(t)}, \mathcal{T}_j\right) \tag{11}$$

Similarly, here $\mathcal{D}_{f_k}(\cdot)$ represents distance metric between two detection responses under representation of feature f_k.

A brief illustration of target representation variation is given in Fig. 4, where three different features are selected for describing a same target in two scenes. It can be observed from Fig. 4 that different features have different reliability in the same scene, and a same feature in different scenes also has different reliability. Our motivation is to select features which have higher reliability for target representation for each given scene, therefore a reliability measurement criterion is proposed and described as follows:

3.3.1. Reliability measurement

In order to explore the feature's reliability for target representation in a given scene, two statistics of variation $v_{f_k}^i(t)$ are studied:

One statistic is the mean of variation μ_{f_k}, formulated as:

$$\mu_{f_k} = \frac{\sum_{l=1}^{t}\sum_{j=1}^{N_r(l)} v_{f_k}^j(l)}{\sum_{l=1}^{t} N_r(l) + 1} \tag{12}$$

where $N_r(t)$ is the number of detection responses associated to tracklets at time t. This statistic μ_{f_k} essentially indicates the average true variation of feature f_k on representing targets in a given scene. Lower true variation means higher stable representation for moving targets.

The other statistic of variation $v_{f_k}^i(t)$ is standard deviation of variation s_{f_k}, formulated as:

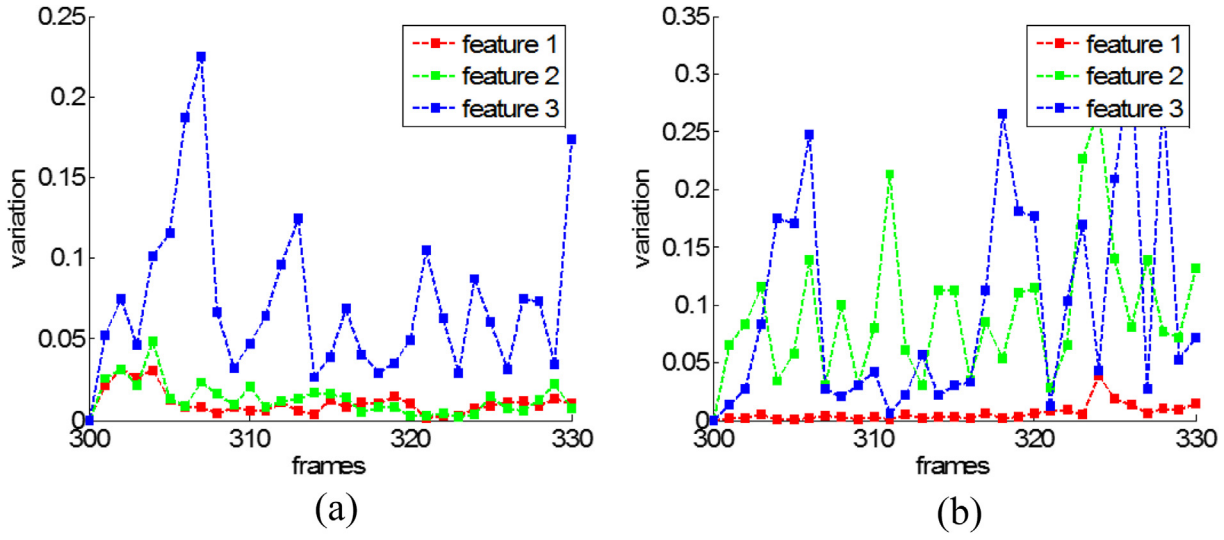


Fig. 4. A brief illustration of feature representation variations in two different scenes, where feature 1–3 are color histogram vectors in HSV (red line), RGB (green line) and YUV (blue line) color space respectively. It can be observed that feature 1 has lower variation in both scenes. Feature 2 has better performance in Scene 1 than Scene 2. Feature 3 has larger variation in both scenes.

$$s_{f_k} = \left(\frac{\sum_{l=1}^{t} \sum_{j=1}^{N_r(l)} \left(v_{f_k}^{j}(l) \right)^2 - \left(\mu_{f_k} \right)^2}{\sum_{l=1}^{t} N_r(l) + 1} \right)^{\frac{1}{2}} \tag{13}$$

where $N_r(t)$ is number of detection responses associated to tracklets at time t. This statistic s_{f_k} essentially indicates stability of feature feature f_k on representing targets in a given scene. Higher observation errors brings higher value of s_{f_k} which means lower stability of the feature representation.

In practice, the ideal feature representation should have small true variation and higher stability, that indicates the variation $v_{f_k}^{j}(t)$ low values of mean μ_{f_k} and standard deviation s_{f_k}.

In this framework, the reliability of feature f_k on target representation can be given as:

$$R_k = U_{f_k} \left(\frac{1}{u_{f_k}} + \omega_r \cdot \frac{1}{s_{f_k}} \right) \tag{14}$$

where U_{f_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. Parameter ω_r is weighting factor to control the impact of μ_{f_k} and s_{f_k} for the reliability R_{f_k}.

3.3.2. Discriminability measurement

In practical multi-tracking, reliable and discriminative feature representation are both vital for data association. Therefore, in the scene-adaptive feature selection, not only the reliability but also the discriminabiity of features should be taken into consideration. For example, in a very crowded scene, even the position feature is very reliable and stable to describe the target, but the distances between them are relatively too small, so the position feature is not a satisfactory feature. Suppose the distance between any two detection responses under representation of feature f_k at time t is given as:

$$v_{f_k}^{ij}(t) = \mathcal{D}_{f_k} \left(\boldsymbol{r}_i^{(t)}, \boldsymbol{r}_j^{(t)} \right) \tag{15}$$

Based on a certain period of observation for a given scene, another statistic d_{f_k} is introduced which is formulated as follows:

$$d_{f_k} = \frac{\sum_{l=1}^{t} \sum_{i=1}^{N_r(l)} \sum_{j=i+1}^{N_r(l)} v_{f_k}^{ij}(l)}{\sum_{l=1}^{t} N_r(l)(N_r(l)-1)/2} \tag{16}$$

where $N_r(t)$ is number of detection responses associated to tracklets at time t. Intrinsically, the statistic d_{f_k} describes the average distance between detection responses in the given scene, under the representation of feature f_k.

Then, the discriminability measurement D_{f_k} is proposed taking two statistics d_{f_k} and μ_{f_k} into consideration, formulated as:

$$D_{f_k} = \frac{1}{U_{f_k}} \left(d_{f_k} - \mu_{f_k} \right) \tag{17}$$

here the parameter U_{f_k} is also used to transform the heterogeneous statistics of different features into homogeneous data to make sure they can be compared together. As mentioned above, μ_{f_k} describes the average true variation of feature f_k on representing same targets in consecutive frames, and d_{f_k} describes the average distance between different targets in the given scene. Naturally, feature with relatively higher d_{f_k} but relatively lower μ_{f_k} is more suitable for a discriminative targets representation in the applied scene. Taking position feature for example, in crowded scenes the mean variation of position feature μ_{pos} is relatively closed to the average position distance d_{pos} compared with less crowded scenes. This indicates that position feature have more discriminative power in less crowded scenes than crowded scenes. Therefore, the discriminability measurement D_k essentially gives a criterion to select features to reduce ambiguity in data association for the scene.

Finally, in practice the reliability measurement R_{f_k} and the discriminability measurement D_{f_k} are combined to calculate the quality of feature f_k for targets representation in the given scene, formulated as follows:

$$Q_{f_k} = \lg R_{f_k} + \omega_q \lg D_{f_k} \tag{18}$$

where ω_q is a weighting factor to control the contribution of reliability and discriminability to the final quality of feature.

4. Depth-invariant part-based appearance model

Besides some low level features mentioned in Section 3.2.3, such as color, motion, etc., appearance is one important high-level feature for target representation, which is widely used in pedestrian-related applications, such as person re-identification [30,39], pedestrian detection [40–43], multi-target tracking [4,29,28], etc. Low-level descriptors are generally used in existing works, but they do not take full advantage of body structure information and result in low discrimination. In this paper, a depth-invariant part-based appearance model is proposed for target representation, which is described as follows:

4.1. Depth invariant transform

First, a depth-invariant transform (DIT) is applied to the original image patch of the detection response, and transform it to a novel depth-invariant image coordinates (DIIC). Suppose the original image coordinates of the detection response is written as (u,v), and the depth-invariant image coordinates is written as $(\widehat{u}, \widehat{v})$, then the depth-invariant transformation can be formulated as:

$$\begin{aligned} \widehat{u} &= a_s \cdot \left(W_o + d \cdot (u - u_0) \cdot \frac{1}{f_u} \right) \\ \widehat{v} &= a_s \cdot \left(H_o + d \cdot (v - v_0) \cdot \frac{1}{f_v} \cdot \cos \alpha_p + d \cdot \sin \alpha_p \right) \end{aligned} \tag{19}$$

where (u_0, v_0, f_u, f_v) is intrinsic parameters of RGB sensor, a_s is the scale factor, and d is the corresponding depth value of coordinate (u,v) on depth image.

Here, W_o is an offset to make sure that after DIT transformation, and the minimum value of $\widehat{u}$ is 1. And H_o is set to make sure the top of all depth invariant patches correspond to a same height in the world coordinates. The parameter α_p in Formula (19) is the pitch angle of the sensor. It can be seen from Fig. 5 that heads of target in different patches are almost at the same horizontal position. An example of appearance modeling procedure is given in Fig. 7. Two detection responses of the same target are transformed to the same scale and same horizontal level. The length metric after DIT transform is proportional to the absolute length metric in the world, and top of two patches both correspond to 1.8 m in the real world.

horizontal ground surface in the applied scenes. Both of them are not general enough because the depth sensor may not capture the data on ground due to smooth floor surface or occlusion, which is a common situation in indoor scenes. Here, we conduct a statistics of detection responses of each target based on a period of observation to the applied scene. A brief illustration of camera pitch angle estimation is given in Fig. 6. For each tracklet, head 3D position coordinates of detection responses belong to this tracklet are collected. After a long period of observation in the given scene, enough reliable data can be collected and the pitch angle of sensor can be estimated. The detailed procedure is given in Algorithm 2.

After the DIT transform, no matter where the target locates in the scene, points with the same height (referred to the ground plane) on the target surface will always correspond to the same horizontal location on the DIIC coordinates, which

Algorithm 2 Automatic Pitch Angle Estimation Algorithm

Input: Tracklets Set $\{\mathcal{T}_1, \cdots, \mathcal{T}_j, \cdots, \mathcal{T}_{N_\mathcal{T}}\}$
Output: Pitch Angle α_p

for $j = 1$ to $N_\mathcal{T}$ **do**
  Get head 3D position coordinates set H_j of detection responses in tracklet $\mathcal{T}_j$:
  $H_j = \{h_j^{(1)}, h_j^{(2)}, \cdots, h_j^{(t_j)}\}$;
  Smoothing head coordinates set H_j to a new set:
  $\hat{H}_j = \{\hat{h}_j^{(1)}, \hat{h}_j^{(2)}, \cdots, \hat{h}_j^{(t_j)}\}$;
  Calculate head motion vectors set:
  $\boldsymbol{v}_j = \left[v_j^1; v_j^2; \cdots; v_j^{t_j-1}\right]$ where $v_j^{t_j} = \hat{h}_j^{(t_j+1)} - \hat{h}_j^{(t_j)}$;
end for
All head motion vector sets are collected to calculate the normal vector $\boldsymbol{n}_g$ of the ground plane:
  $\boldsymbol{V} \cdot \boldsymbol{n}_g = \boldsymbol{0}$ where $\boldsymbol{V} = \left[\boldsymbol{v}_1; \boldsymbol{v}_2; \cdots; \boldsymbol{v}_{N_\mathcal{T}}\right]$
Adopting RANSAC algorithm to calculate the normal vector $\boldsymbol{n}_g$:
  $\boldsymbol{n}_g = \arg\min_{\boldsymbol{n}_g \in \mathbb{R}^3} \|\boldsymbol{n}_g \cdot \boldsymbol{V}\|$
Given the unit vector $\boldsymbol{u}_d$ along depth axis, the pitch angle α_p can be calculated as:
  $\alpha_p = \arccos < \boldsymbol{n}_g \cdot \boldsymbol{u}_d >$

It is worth to mention that the parameter α_p in Formula (19) is the pitch angle of the sensor. In order to make the DIT transform works fine with different camera pitch angles in various scenes without manually calibration, we proposed an automatic pitch-angle estimation method which are suitable for depth data, which is given in Algorithm 2. Previous works like [52,51] usually calculate the normal vector of the ground plane and use it to estimate camera pitch angle. For example, work in [51] manually selects a portion of depth image area that corresponds to the ground plane. Although work in [52] is an automatic method to determine the ground plane based on 3D points cloud, but it is based on a strong assumption that many vertical walls exist as well as a large and roughly

makes a firm foundation for the following combination of spatial information and basic features. Therefore, the DIT transform makes the appearance model invariant to scale change and view-truncation on 2D image plane.

4.2. Part-based modeling

The spatial layout information is is considerably critical information for appearance modeling and human targets is not a rigid object for its complex kinematics, so it can be better described using a part-based model [53]. After the DIT transform, based on the detection response in the novel depth-invariant image coordinates, we combine basic features

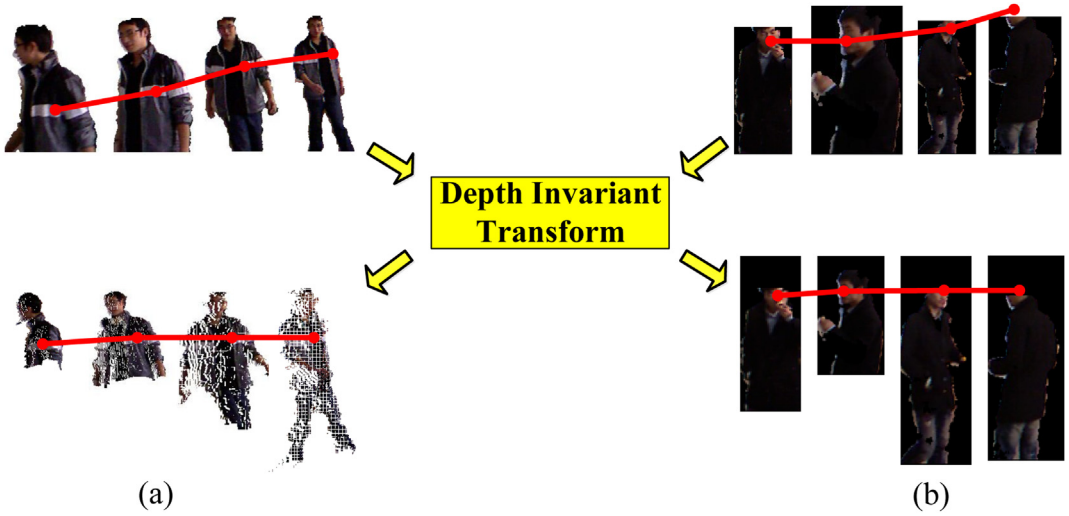


Fig. 5. The figure shows patches of the two targets' detection responses before and after DIT during the tracking process. Two versions of image patches with black and white backgrounds are shown to give a better visualization of the transformed patches. The red lines indicate same body parts in the real world correspond to same part on the image patch after DIT.

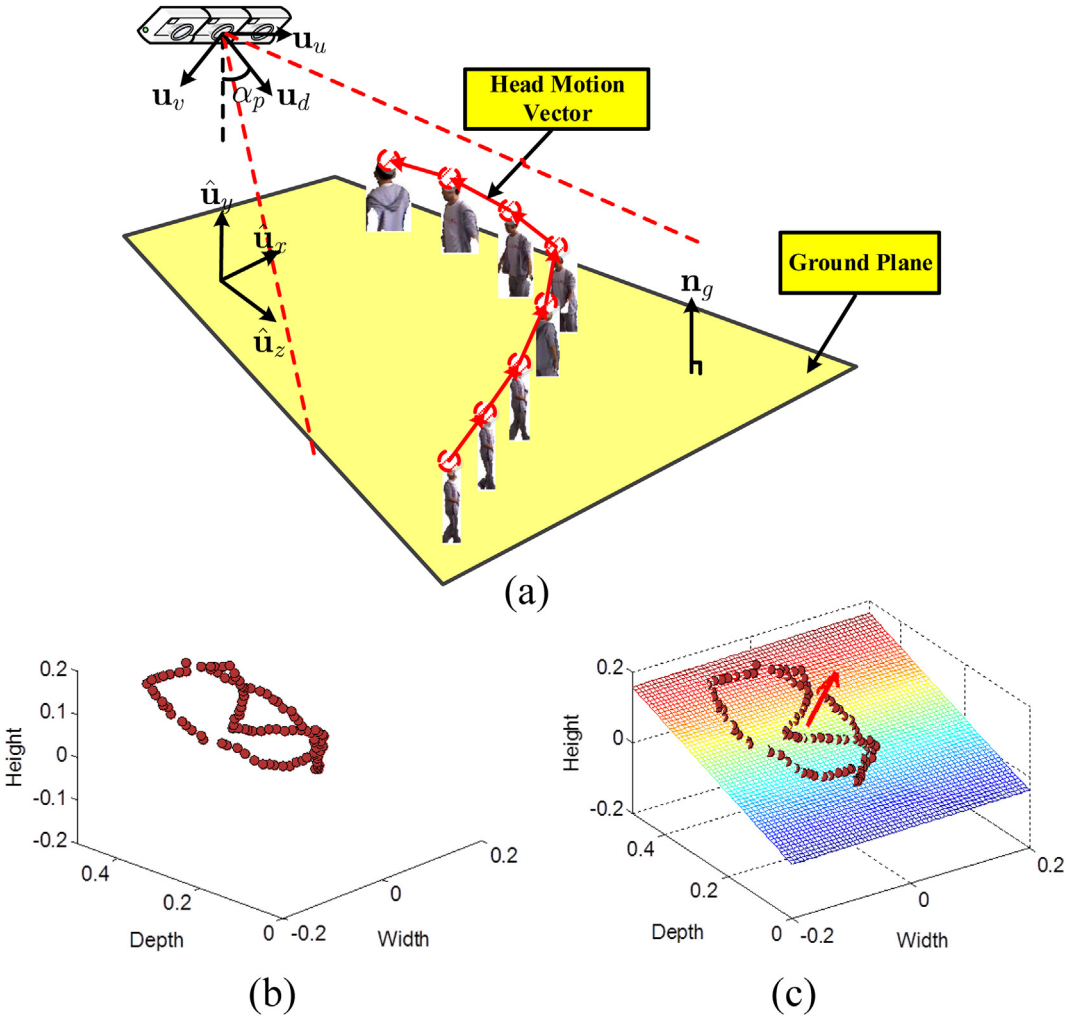


Fig. 6. A brief illustration of camera pitch angle estimation through long-term observation in the applied scene.

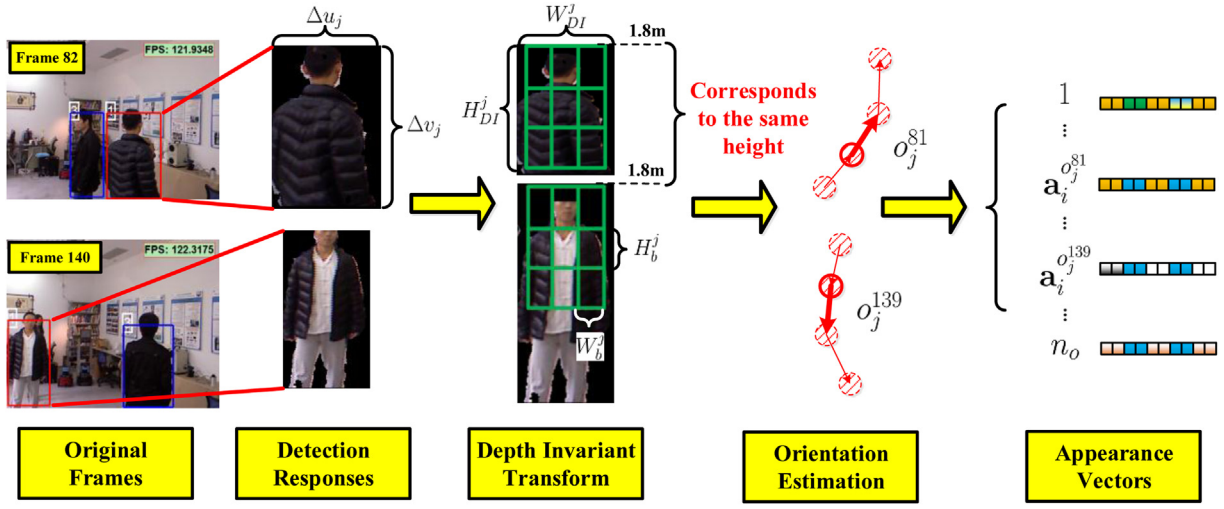


Fig. 7. Scale of the red target's detection responses changes in 2D image plane when target moves near and far, but it maintains relatively stable scale after Depth Invariant Transform (DIT).

according to part-based spatial layout to complete the appearance modeling.

For basic features selection, several most used features are tested and selected in our framework. Same as the previous work [53], several low-level features are used for appearance modeling, including weighted HSV (wHSV) color histograms, SIFT descriptors and LAB color histograms, where weighted wHSV color histograms are extracted to capture color information as suggested in [54], SIFT descriptors are used to capture texture information and handle illumination variation and LAB color histograms are extracted to enhance illumination invariance.

Different from pervious part-based approaches, which usually divide the image patch of detection response into several part blocks in 2D image plane, In this work the block size of each part has a depth-invariant size (W_b^j, H_b^j), which correspond to fixed values in the world coordinates, as shown in Fig. 7, This makes the appearance model tolerable to large scale changes. In order to avoid bad effects of view-truncations, as bottom-half body is often truncated by filed-of-view, only the top-half of a target is modeled.

For each part, one or more basic features are used for appearance vector construction, according to the task requirements of the applied scene. Assume the appearance vector of a part block is denoted as $\boldsymbol{a}_{b_i}$, where b_i is the index of the blocks, the part-based model of each detection response $\boldsymbol{r}_j$ can be 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 appearance vector of a part block and bn is number of part blocks.

Thus, given two constructed appearance vector $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$, appearance affinity between $\mathcal{T}_j$ and $\boldsymbol{r}_i$ can also be given as the following general form:

$$\mathcal{A}_{\boldsymbol{a}}(\boldsymbol{r}_i|\mathcal{T}_j) = G\left(\frac{1}{\mathrm{corr}(\boldsymbol{a}_i, \boldsymbol{a}_j)}; \mu_{\boldsymbol{a}}, \Sigma_{\boldsymbol{a}}\right) \tag{20}$$

In this framework, similar to the related work in [14], features' affinities are all defined by Gaussian distributions. Variances μ_{f_k} and Σ_{f_k} are given based on statistics mean μ_{f_k} and standard division s_{f_k} during long term observation following the algorithm proposed in Section 3.3.

4.3. Orientation-based modeling

In indoor environments, appearance of a same target may vary a lot due to turning around, as shown in Fig. 7, the back and front appearances of the target vary a lot. In order to address this situation, orientation of target is utilized to guide appearance model in this work. For a tracklet $\mathcal{T}_j$, the coarse orientation of its response $\boldsymbol{t}_j^{t-1}$ at time $t-1$ is given by:

$$o_j^{t-1} = \sigma_{n_o}\left(\arctan\frac{Zc_j^{t-1} - Zc_j^{t-2}}{Xc_j^{t-1} - Xc_j^{t-2}}\right) \in [1 \cdots n_o] \tag{21}$$

where σ_n is a quantization function which quantify the orientation of a response into n_o levels. Hence the appearance $\boldsymbol{a}_j$ and color model $\boldsymbol{c}_j$ of a tracklet $\mathcal{T}_j$ are divided into n_o models $\{\boldsymbol{a}_j^1, \cdots, \boldsymbol{a}_j^{n_o}\}$ and $\{\boldsymbol{c}_j^1, \cdots, \boldsymbol{c}_j^{n_o}\}$.

Thus, considering orientation of the detection response, its color affinity and appearance affinity can be reformulated as follows:

$$\mathcal{A}_{\boldsymbol{c}}(\boldsymbol{r}_i|\mathcal{T}_j) = G\left(\frac{1}{\mathrm{corr}\left(\boldsymbol{c}_i^{o_i}, \boldsymbol{c}_j^{o_j^{t-1}}\right)}; \mu_{\boldsymbol{a}}, \Sigma_{\boldsymbol{c}}\right) \tag{22}$$

$$\mathcal{A}_{\boldsymbol{a}}(\boldsymbol{r}_i|\mathcal{T}_j) = G\left(\frac{1}{\mathrm{corr}\left(\boldsymbol{a}_i^{o_i}, \boldsymbol{a}_j^{o_j^{t-1}}\right)}; \mu_{\boldsymbol{a}}, \Sigma_{\boldsymbol{a}}\right) \tag{23}$$

where $corr(\boldsymbol{v}_i, \boldsymbol{v}_j)$ calculates the correlation between vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$.

In practice application, it is also a scene-adaptive problem whether adopting multiple or single orientation appearance model. For example, in some scenes targets always move towards regular orientations, which means that the orientation of

the target seldom changes or changes smoothly between consecutive frames when it moves in the filed of view, so the multi-orientation scheme is less necessary in such kind of scenes. On the contrary, in some scenes targets may wander in circles or exhibit long term occlusion, their detection responses always have multiple orientations.

5. Experiment and analysis

In this section, we conducted several experiments for observation, comparison and analysis from the following three sections to verify the effectiveness of the proposed framework: First, we verify the necessity and effectiveness of scene-adaptive feature selection framework in Section 5.2, based on analyzing the reliability differences of features on various scenes, and compare the proposed scene-adaptive hierarchical data association framework with non-scene-adaptive framework in various scenes on both RGB-D datasets and RGB datasets, to evaluate the generality of the proposed framework. Then, we conduct a series of experiments to evaluate performance of the depth invariant part-based appearance model (DIAM) in various scenes, which is given in Section 5.3. Third, we analyze advantages and disadvantages of the proposed hierarchical data association framework and non-hierarchical framework, conducting experiments on RGB-D datasets and combining various features, which is elaborated in Section 5.4.

5.1. Datasets and settings

There is no generally accepted benchmark available for multi-tracking [5]. 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.

Our experiments are mainly based on the RGB-D datasets which can evaluate the whole proposed framework, such as the depth-invariant appearance model, the pitch angle estimation module and the scene-adaptive feature selection module which includes many 3D features. All these modules need both RGB and depth data. For the RGB datasets, we only use it to evaluate the performance of the proposed hierarchical data association framework, in order to make a detailed comparison with classic framework in this filed. For RGB datasets, most related publications have carried out experiments on their own sequences. Besides our own RGB datasets, we combine several of them which are public into our RGB datasets for comparison, such as PETS09 S2.L1 – S2.L3 [49], TUD Campus [50], TUD Crossing [50]. The detailed configuration of the experimental datasets is given in Table 1.

Compared with long period run of an online multi-tracking system, feature selection is a short period procedure, which does not need to be 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_{f_k} for all features in feature space are calculated according to algorithm given in Section 3.3. Due to its statistical nature, occasional id-switches can be overlooked in data association. The variation reference U_{f_k} in Formula (14) can be obtained by averaging variations of each feature f_k during association in several sequences in different scenes. It is a constant value based on prior statistics which is scene-independent. The weighting factor ω_r is empirically set to 0.5 to control the impact of μ_{f_k} and s_{f_k} for the reliability R_{f_k} in Formula (14).

Table 1
Detailed configuration of experimental datasets.

Dataset name	Dataset type			
	In/Out	Distance	Crowdedness	RGB-D/RGB
Scene 1 – Scene 5	Indoors	Close	Crowded	RGB-D
Scene 6 – Scene 10	Indoors	Close	Crowded	RGB
Scene 11 – Scene 15	Outdoors	Distant	Less Crowded	RGB
PETS09 S2.L1 – S2.L3 [49]	Outdoors	Distant	Less Crowded	RGB
TUD campus [50]	Outdoors	Close	Crowded	RGB
TUD crossing [50]	Outdoors	Close	Crowded	RGB

For our approach, α and a_s is set to 10 in Formulas (7) and (19). H_{off} is set to 100 in Formula (19). For part-based appearance model, parts number bn is set to 9 and block size $\{H_b^j, W_b^j\}$ (shown in Fig. 7) 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 (see Fig. 8).

5.2. Scene-adaptive feature selection evaluation

First, we verify the necessity and effectiveness of scene-adaptive feature selection framework based on analyzing the reliability differences of features on various scenes.

A standard metric for evaluating object trackers is the Multiple Object Tracking Accuracy (MOTA) [48], defined as

$$MOTA = 1 - \frac{\sum_t \left(c_m(m_t) + c_f(fp_t + c_s(mme_t))\right)}{\sum_t g_t} \tag{24}$$

where g_t is the number of ground truth detections, m_t the number of miss-detections, fp_t is the false positive count and mme_t the number of instantaneous identity switches [12]. Here, we adopt MOTA to to evaluate performance of multi-tracking performance. In addition, taking into consideration of evaluating real-time capability of algorithm, we also compute the frames per second (FPS).

The quantitative results are given in Table 4. Here non-HDA stands for combining all features in feature space for data association. It can be seen from Table 2 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

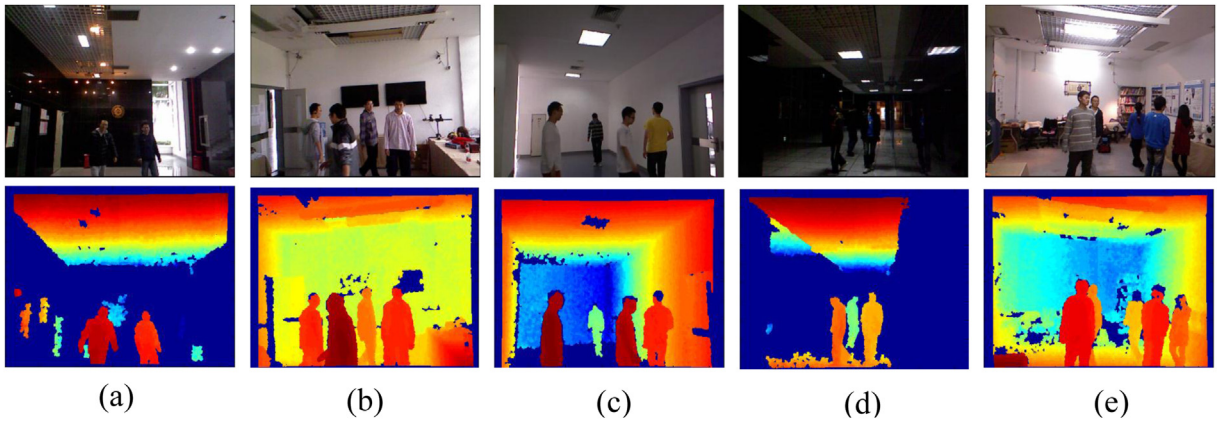


Fig. 8. Figures from (a) to (e) are samples of our RGB-D datasets recorded in Scene 1 to Scene 5.

Table 2
Quantitative accuracy and speed comparison (not including detection time, only data association) of several data association schemes in our datasets (Table 1).

Frameworks	MOTA (%)/FPS			
	Scene 1–5	Scene 6–10	Scene 11–15	Public
SAFS + HDA	**91.20/29.3**	**86.23/20.5**	**89.20**/23.4	**93.25/21.9**
HDA	87.65/25.6	82.67/19.5	86.32/**23.6**	90.95/20.1
non-HDA	75.25/16.3	71.36/15.5	73.95/18.6	88.24/19.2

The significance of bold characters indicates the best experimental results or performance of the corresponding experiments group.

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 2 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.

5.3. Depth-invariant part-based appearance model evaluation

In this section we conduct a series of experiments to evaluate performance of the depth invariant part-based appearance model in various scenes. Comparative experiments are conducted with classic average-division part-based appearance model (ADPAM) [28], and a state-of-the-art classifier-based discriminative appearance model (CDAM) [4], which verifies the overall performance of the depth-invariant part-based appearance model (DIPAM).

First, we evaluate robustness of our depth-invariant part-based appearance model (DIPAM) in the RGB-D datasets. Comparative experiments are conducted with two most representative appearance modeling schemes in the multi-tracking field, one is classic average-division part-based appearance model (ADPAM) which divides response's bounding box into blocks averagely which is used in [28], and the other is a state-of-the-art classifier-based discriminative appearance model (CDAM) proposed in [4] which combines color histograms, covariance matrixes, and histogram of gradients (HOG) for appearance modeling. In order to achieve a pure comparison of appearance models, all these three appearance models are all fused in the proposed hierarchical data association framework with fixed hierarchical feature space. The fixed hierarchical feature space is constructed with four widely used features: position, motion, color and appearance models. Thus, based on this fixed hierarchical feature space construction, comparative experiments with DIPAM, ADPAM and CDAM are conducted on different scenes, and the quantitative experimental results are given in form of MOTA (%) and FPS, shown in Table 3. It can be seen from the data that DIPAM has relative higher accuracy (MOTA) compared with the other two methods. ADPAM has low MOTA because it only divides the 2D image patch of the detection response averagely into part blocks, so it is vulnerable to large size variation and truncation by the field-of-view. Take detection responses shown in Fig. 5 for example, the detection responses always exhibit half-body or whole body due to the truncation, it is no longer suitable for these classic appearance modeling methods such as ADPAM. The CDAM performs well in MOTA because it combines bunch of local descriptors for appearance modeling and its classifier-based nature, but is more computation-consuming in appearance modeling with calculation of many local descriptors and classifier updating. In conclusion, DIPAM is an appearance modeling scheme based on DIT transformation, which makes the image patch content in a part-block corresponds to its real location on the real world. This makes it robust to scale variation and view-truncation which commonly occur in

Table 3
Performance of our depth-invariant part-based appearance model and comparative appearance models in different scenes.

MOTA (%)/FPS	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5
DIPAM (ours)	**88.5**/21.3	**82.0**/20.4	**79.5**/21.5	**86.3**/22.7	**79.5**/18.2
ADPAM [28]	76.5/20.2	70.2/22.5	65.2/23.0	78.2/21.5	61.5/19.5
CDAM [4]	87.5/10.7	81.5/11.6	73.2/12.3	82.4/15.0	75.2/10.2

The significance of bold characters indicates the best experimental results or performance of the corresponding experiments group.

indoor scenes, and improves its description ability for appearance modeling.

5.4. Hierarchical data association evaluation

Besides the whole performance evaluation given in Section 5.2, in this section we will conduct more detailed experiments to evaluate the effectiveness of the HDA scheme. We organize the experiments from two aspects: First, we analyze how hierarchical feature space construction influence the performance of HDA. Second, we compare hierarchical data association (HDA) scheme with non-hierarchical data association (non-HDA) scheme while applying the same hierarchical feature space.

As shown in the top part of Table 4, various classic features for target representation are used for feature space construction, including position (P), size (S), motion (M), color (C) and appearance model (A), which have been elaborated in Section 3.2.3 and Section 4. Here, HDA data association scheme works on self-constructed hierarchical feature space with six different feature permutations (PSMC, PSM, ACPM, ACP, PMCA, PMC). Quantitative results in form of MOTA (%) and FPS are given in Table 4.

It can be seen from Table 4 that different feature permutations contribute to different performance of HDA in the same scene. On the other hand, the same feature permutation exhibits different performance in different scenes. For example, Scene 1 is a less-crowded room with poor illumination conditions, so hierarchical feature space constructed with position, size and motion (PSM) permutation exhibits better performance. On the contrary, feature permutations such as ACPM and PMCA which are constructed with appearance and color cues, perform badly in this scene. This is because appearance and color cues have relatively lower reliability in this scene due to base illumination conditions. But appearance and color cues play important roles in some crowded scenes with better illumination, such as Scene 3 and Scene 5. This indicates that with the same features set, such as ACPM and PMCA, different feature permutations will lead to different HDA performances, and even the same feature permutation can lead to different HDA performances in different scenes. Therefore, scene-adaptive feature selection and feature space construction are vital for the following HDA in real applications in various scenes.

Table 4
Quantitative accuracy and speed comparison (not including detection time, only data association) between HDA and non-HDA data association schemes with various feature permutations in our datasets.

Scheme	MOTA (%)/FPS					
	Features	Scene 1	Scene 2	Scene 3	Scene 4	Scene 5
HDA	PSMC	78.3/18.2	82.6/16.8	80.3/15.4	80.2/15.6	77.3/15.2
	PSM	**82.3/24.6**	78.5/**23.2**	74.2/**24.9**	**81.5/21.2**	72.8/**22.9**
	ACPM	70.2/19.5	80.6/16.4	77.2/14.3	65.3/12.7	82.6/14.3
	ACP	65.4/17.0	78.3/13.4	74.5/12.8	69.8/13.5	81.5/14.6
	PMCA	78.3/20.6	**86.4**/18.3	**85.5**/17.5	76.7/19.4	**86.5**/15.3
	PMC	80.3/19.2	85.2/19.5	82.7/18.4	75.3/18.5	84.4/17.2
non-HDA	PSMC	65.2/12.3	71.3/11.5	73.9/12.6	62.2/13.2	70.2/10.8
	PSM	73.2/15.2	**68.2**/15.3	**70.7**/14.5	75.6/14.9	**67.3**/13.5
	ACPM	**65.7/6.3**	72.3/**7.5**	73.9/**9.6**	**60.2/7.2**	69.2/**6.8**
	ACP	66.8/8.5	74.5/9.8	73.2/12.9	61.4/7.9	72.9/9.9
	PMC	69.3/13.3	71.6/14.5	72.5/13.2	65.1/13.8	67.2/12.8

The significance of bold characters indicates the best experimental results or performance of the corresponding experiments group.

Then we evaluate advantages of the hierarchical data association scheme (HDA) to the non-hierarchical data association scheme (non-HDA). Quantitative results in form of MOTA (%) and FPS are given in the bottom part of the Table 4. Compared with the HDA results in Table 4, it is easy to observe that the non-HDA scheme is relatively poor in terms of computing speed, which can be seen from its lower FPS values compared with the HDA scheme. This is because the non-HDA scheme always combines all selected features and calculates their target representations in every data association. On the contrary, the HDA scheme gradually combines features in the hierarchical feature space according to the need of discriminating ambiguous detection responses in the current data association task, which leads to lower computation cost. On the other hand, the accuracy (MOTA) of non-HDA scheme is relatively lower than the HDA scheme.

6. Conclusions

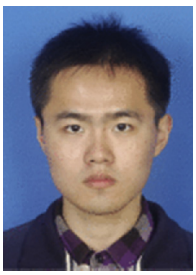
In this work, we focus on efficiently combing features to discriminate ambiguous targets for better data association, and handling large appearance variations in indoor environments. Compared with previous work, the proposed hierarchical data association scheme based on hierarchical feature space gradually fuses more features according to requirements of distinguishing conflicting responses, leading to less error accumulation and least computational cost. Moreover, scene-adaptive thinking is introduced to our framework and 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, which increase applicability and generality of the framework. 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 method demonstrates good performance in various challenging indoor scenes running in real-time. Experimental results demonstrate 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 scheme in various scenes, with the scene-adaptive thinking, and will pay more efforts on more discriminative target representation for better data association.

References

[1] C. Lu, C. Wu, L. Fu, IEEE Trans. Syst. Man Cybern. Part C TSMC 41 (1) (Jan. 2011) 120–129.
[2] P. Rashidi, D.J. Cook, L.B. Holder, S.E. Maureen, IEEE Trans. Knowl. Data Eng. TKDE 23 (4) (Apr. 2011) 527–539.
[3] J. Han, E.J. Pauwels, P.M. de Zeeuw, P.H.N. de With, IEEE Trans. Consumer Electron. 58 (2) (May 2012) 253–263.

[4] C. Kuo, C. Huang, R. Nevatia, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2010) 685–692.
[5] M.D. Breitenstein, F. Reichlin, B. Leibe, E.K. Meier, L.V. Gool, IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 33 (9) (Sept. 2005) 1820–1832.
[6] C.X. Liu, S.G. Gong, C.C. Loy, X.G. Lin, Eur. Conf. Comput. Vis. Work. ECCVW (2012) 391–401.
[7] B. Prosser, W. Zheng, S. Gong, T. Xiang, Br. Mach. Vis. Conf. BMVC 1 (3) (2010) 1–11.
[8] W.S. Zheng, S.G. Gong, T. Xiang, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2011) 649–656.
[9] J. Berclaz, F. Fleuret, P. Fua, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2006) 744–750.
[10] J. Xing, H. Ai, S. Lao, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2009) pp.1200–1207.
[11] L. Zhang, Y. Li, R. Nevatia, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2008) 1–8.
[12] J. Berclaz, F. Fleuret, E. Turetken, P. Fua, IEEE Int. Conf. Comput. Vis. ICCV (2011) 137–144.
[13] J. Berclaz, F. Fleuret, E. Turetken, P. Fua, IEEE Trans. Pattern Analy. Mach. Intell. TPAMI 33 (9) (Sept. 2011) 1806–1819.
[14] C. Huang, B. Wu, R. Nevatia, Eur. Conf. Comput. Vis. ECCV (2008) 788–801.
[15] M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L.V. Gool, IEEE Int. Conf. Comput. Vis. ICCV (2009) 1515–1522.
[16] Y. Cai, N. de Freitas, J.J. Little, Eur. Conf. Comput. Vis. ECCV (2006) 107–118.
[17] K. Okuma, A. Taleghani, O.D. Freitas, J.J. Little, D.G. Lowe, Eur. Conf. Comput. Vis. ECCV (2004) 28–39.
[18] B. Wu, R. Nevatia, Int. J. Comput. Vis. IJCV 75 (2) (Nov. 2007) 247–266.
[19] D.R. Magee, Image Vis. Comput. (2004) 143–155.
[20] Z. Khan, T. Balch, F. Dellaert, IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 27 (11) (Nov. 2005) 1805–1891.
[21] W. Choi, C. Pantofaru, S. Savarese, IEEE Int. Conf. Comput. Vis. Work. ICCVW (2011) 1076–1083.
[22] M. Yang, F. Lv, W. Xu, Y. Gong, IEEE Int. Conf. Comput. Vis. ICCV (2009) 1554–1561.
[23] P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, IEEE Trans. Pattern Anal. Mach. Intell. TPAMI 32 (9) (2010) 1627–1645.
[24] M. Hofmann, M. Haag, G. Rigoll, IEEE Int. Workshop Perform. Eval. Track. Surveillance PETS (2013) 22–28.
[25] Hoan Nguyen, T. Fasciano, D. Charbonneau, A. Dornhaus, M.C. Shin, IEEE Winter Conf. Appl. Comput. Vis. WACV (2014) 941–946.
[26] M. Baum, P. Willett, IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP (2014) 4209–4213.
[27] B. Yang, R. Nevatia, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2012) 2034–2041.
[28] B. Yang, R. Nevatia, Eur. Conf. Comput. Vis. ECCV (2012) 484–498.
[29] B. Yang, R. Nevatia, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2012) 1918–1925.
[30] L. Bazzani, M. Cristani, V. Murino, Comput. Vis. Image Underst. CVIU 117 (2) (2013) 130–144.
[31] N. Bird, O. Masoud, N. Papanikolopoulos, A. Isaacs, IEEE Trans. Intelligent Transp. Syst. TITS 6 (2) (June 2005) 167–177.
[32] C. Wang, H. Liu, L. Ma, IEEE Signal Process. Lett. SPL 21 (6) (Jun. 2014) 717–721.
[33] B. Ni, N.C. Dat, P. Moulin, IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP (2012) 1405–1408.
[34] J. Shotton, et al., Commun. ACM (2013) 116–124.
[35] L. Spinello, K.O. Arras, IEEE Int. Conf. Robotics Automation ICRA (2012) 4469–4474.
[36] L. Spinello, K.O. Arras, IEEE/RSJ Int. Conf. Intelligent Robots Syst. IROS (2011) 3838–3843.
[37] L. Cruz, Patterns Images Tutorials SIBGRAPI (2012) pp.36–49.
[38] F. Sanchez, L. Rubio, J. Diaz, E. Ros, Mach. Vis. Appl. MVA 25 (5) (Oct. 2013) 1211–1225.
[39] C. Liu, S. Gong, C.C. Loy, Pattern Recognit. 47 (4) (2014) 1602–1615.
[40] J. Marin, D. Vazquez, D. Geronimo, A.M. Lopez, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2010) 137–144.
[41] M. Liem, D.M. Gavrila, IEEE Int. Conf. Work. Automatic Face Gesture Recognit. (2013) 1–6.
[42] P. Sermanet, K. Kavukcuoglu, S. Chintala, Y. LeCun, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2013) 3626–3633.
[43] M. Tang, Andriluka, B. Schiele, Int. J. Comput. Vis. IJCV (2014).
[44] S. Bak, E. Corvee, F. Bremond, M. Thonnat, IEEE Int. Conf. Adv. Video Signal Based Surveillance AVSS (2010) 1–8.
[45] L. Bazzani, M. Cristani, A. Perina, V. Murino, Pattern Recognit. Lett. PRL 33 (7) (May 2012) 898–903.
[46] Y. Lu, L. Lin, W.S. Zheng, IEEE Int. Conf. Comput. Vis. CVPR (2013) 3152–3159.
[47] H. Liu, Z. Yu, H.B. Zha, L. Zhang, Y.X. Zou, Pattern Recognit. Lett. PRL 30 (9) (July 2009) 827–837.
[48] K. Bernardin, R. Stiefelhagen, EURASIP J. Image Video Process. JIVP 2008 (1) (Jan. 2008) 1–10.
[49] J. Ferryman, IEEE Workshop Perform. Eval. Track. Surveillance (2009).
[50] M. Andriluka, S. Roth, B. Schiele, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR (2008) 1515–1522.
[51] A. Albiol, J. Oliver, J.M. Mossi, IET Comput. Vis. 6 (5) (Nov. 2012) 378–387.
[52] W. von Hansen, Photogramm. Image Anal. PIA (2007) 93–97.
[53] H. Liu, M. Qian, C. Wang, IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP (2015).
[54] M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) (2010) 2360–2367.

Prof. Hong Liu received the Ph.D. degree in mechanical electronics and automation in 1996, and serves as a Full Professor in the School of EE&CS, Peking University (PKU), China. Prof. Liu has been selected as Chinese Innovation Leading Talent supported by "National High-level Talents Special Support Plan" since 2013. He is also the Director of Open Lab on Human Robot Interaction, PKU, his research fields include computer vision and robotics, image processing, and pattern recognition. Dr. Liu has published more than 150 papers and gained Chinese National Aero-space Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Professors in PKU. He is an IEEE member, vice president of Chinese Association for Artificial Intelligent (CAAI), and vice chair of Intelligent Robotics Society of CAAI. He has served as keynote speakers, co-chairs, session chairs, or PC members of many important international conferences, such as IEEE/RSJ IROS, IEEE ROBIO, IEEE SMC and IIHMSP, recently also serves as reviewers for many international journals such as Pattern Recognition, IEEE Trans. on Signal Processing, and IEEE Trans. on PAMI.



Yuan Gao received the B.E. degree in intelligent science and technology from Xidian University in 2012. Then he obtained the M.S. degree in computer applied technology from Peking University in 2015. Currently, he is working toward the Doctor degree in Christian-Albrechts-University of Kiel, Germany. His research interests include object detection, 3D reconstruction, facial expression and gender recognition. He has published articles in IEEE International Conference on Image Processing (ICIP).