Available online at [www.sciencedirect.com](http://www.sciencedirect.com/science/journal/24682322)

ScienceDirect

[CAAI Transactions on Intelligence Technology 1 (2016) 210e224](http://dx.doi.org/10.1016/j.trit.2016.10.008)

<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. Copyright © 2016, Chongqing University of Technology. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

*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](#_bookmark24)e[3]](#_bookmark24), issues like multi-tracking [[4,5]](#_bookmark25), and re-identification [[6](#_bookmark27)e[8]](#_bookmark27) attract more and more attention from

\* This work is supported by National Natural Science Foundation of China (NSFC, No. 61340046), National High Technology Research and Develop- ment 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](mailto:hongliu@pku.edu.cn) (H. Liu), [canwang@pku.edu.cn](mailto:canwang@pku.edu.cn) (C. Wang), [ygao@sz.pku.edu.cn](mailto:ygao@sz.pku.edu.cn) (Y. Gao).

Peer review under responsibility of Chongqing University of Technology.

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]](#_bookmark25). However, this is highly challenging in crowd environments with frequent oc- clusion, 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](#_bookmark28)e[13]](#_bookmark28) to get better association via global analysis, like global trajectory optimization [[9]](#_bookmark28), network flows [[11]](#_bookmark29), hierarchical tracklets association [[14]](#_bookmark31), etc. However, they are based on the prereq- uisite that detection responses in all frames are given, both

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

2468-2322/Copyright © 2016, Chongqing University of Technology. Production and hosting by Elsevier B.V. This is an open access article under the CC BY-NC- ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

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 de- cisions [[15](#_bookmark32)e[18]](#_bookmark32). They usually relies on Kalman [[19]](#_bookmark33) or par- ticle filter [[20]](#_bookmark34) to handle data association. Because of their recursive nature, this category is suitable for time-critical ap- plications, 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]](#_bookmark25) 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 chal- lenging 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, obser- vation 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](#_bookmark0)), espe- cially 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 perfor- mance 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 discrimi- nating 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 hi- erarchical 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 associ- ation. 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 envi- ronments. 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.

1. Related work
   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]](#_bookmark35).

detection and pedestrian detection [[23]](#_bookmark36), which links current detection responses to existing trajectories during multi- tracking. Recently, more and more works adopt data associa- tion based tracking scheme to implement robust multi-tracking [[24](#_bookmark37)e[26]](#_bookmark37) while utilizing different frameworks. The data asso- ciation 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]](#_bookmark28) and network flows [[11]](#_bookmark29). 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.

* 1. *Feature representation and appearance models*

Feature representation of targets is crucial in data associa- tion 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]](#_bookmark31), like color, size, motion, position, etc., and our framework also follows this scheme for better prac- tical 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]](#_bookmark25) 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]](#_bookmark40) equally subdivides the bounding box of detection response into ten horizontal stripes, and work in [[28]](#_bookmark38) 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

appearance model based on RGB-D data, which will be described in details in Section [4](#_bookmark11).

* 1. *RGB-D data for indoors applications*

Recently, combining RGB and depth data for computer vision applications becomes more and more popular [[37]](#_bookmark45), because recent emergence of depth sensor (e.g., Microsoft Kinect) has made it feasible and economically sound to cap- ture in real-time not only color images but also depth maps with appropriate resolution and accuracy [[33]](#_bookmark42). 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 com- plementary 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]](#_bookmark41), background subtraction [[38]](#_bookmark46) and 3D body pose estimation [[34]](#_bookmark43), which provides robust and various target detection schemes. Many depth-based features are also proposed by previous works [[35,36]](#_bookmark44) 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 condi- tions and scale variation on 2D image plane.

1. Scene-adaptive hierarchical data association
   1. *Preliminaries*

In the multi-cue data association framework, the key prob- lem is to associate *n* detection responses in the current frame with *m* existing tracklets. Through out this paper, let ℛ*t* := {*ri*}*n* denote *n* detection responses at frame *t* and let *ri*

denote one detection response. Let У := {У *j*}*m* denote *m*

existing tracklets and let У *j* denote one tracklet, formulated as:

У := n/; *r*(*t*—2); *r*(*t*—1)o (1)

*j j j*

include meaningful features. In addition to these simple and intuitive spatial layouts, work in [[30]](#_bookmark39) 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 oc- clusion, caused by unsatisfied detector performance in complicated environments. Therefore, in order to handle this problem, we propose a novel depth-invariant part-based

here *r*(*t*—1) denotes the detection response associated to tracklet У *j* at time *t*—1. It can be seen that one tracklet actually con- tains all detection responses associated to it in the past. At time *t* during data association, for each new detection response *r*(*t*), we should associate it to its corresponding tracklet У *j*.

The most commonly used approach is to calculate affinities between a detection response *ri* and a tracklet У *j* under rep- resentation of one or more features, and then to multiply all affinities to obtain the final association probability.

*j*

*i*

In classic association frameworks [[4,14,28]](#_bookmark25), link probabil- ity between *ri* and У *j* is usually defined as the product of af- finities based on several features, like position, size, appearance, etc., formulated as:

*Plink ri*; У *j* = A*pos ri*; У *j* A*sz ri*; У *j* A*ap ri*; У *j* / (2)

where A(*ri*; У *j*) denotes the affinity between a detection response *ri* and a tracklet У *j*, and subscribes *‘pos*’, *‘sz*’ and *‘ap*’ denotes position, size and appearance features used for target representation respectively. For each association, pre- vious 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 represen- tations and brings unnecessary computational cost.

A brief example of how observation errors affect data as- sociation is given in [Fig. 2](#_bookmark2). Let У 1 and У 2 denote true values of two tracklets in a given feature space, in other words under representation of a given feature *fk*. Because any feature rep- resentation can be regarded as a data point in its own feature space. Let У 1 and У 2 denote their observed values. Here *ri* and *rl* denote two detection responses detected in the current frame, and *ri* and *rl* are their corresponding true values. It can be seen that the observed affinities between У 1 and two re- sponses *ri* and *rl* are almost the same, but the true values differ

a lot. This is due to observation errors *eri* , *erl* and *e*У 1 between

*f f f*

*k k k*

observed detection responses and their true positions in the

given feature space.

Unfortunately, this is the case for almost all feature repre- sentations in practical application. This indicates that if even non-trivial observation errors exist under several feature rep- resentation, they will accumulate when multiplying feature affinities, as the classic works did in [Formula (2)](#_bookmark1). Therefore, the final link probability *Plink* in [Formula (2)](#_bookmark1) may not reflect the real affinity between detection responses and tracklets. Even there exists reliable and discriminating feature repre- sentations 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 as- sociation scheme with reliable feature representation, which is the main focus in this work.

* 1. *Hierarchical data association*

As mentioned above, combining a bunch of features may not contribute to a better association between detection re- sponses 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 illu- mination, which are common situations in indoor scenes, so this may bring larger observation errors to targets represen- tation; 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 hierar- chical data association scheme is proposed as follows:

* + 1. *Hierarchical feature space construction*

First, a feature space 7 is constructed with various com- mon used features {*fk*} for describing detection responses. Based on the feature space 7 , a generative form of the link probability is formulated as:

*Plink* *ri*; У *j*|7 = Y A*fk* *ri*; У *j* (3)

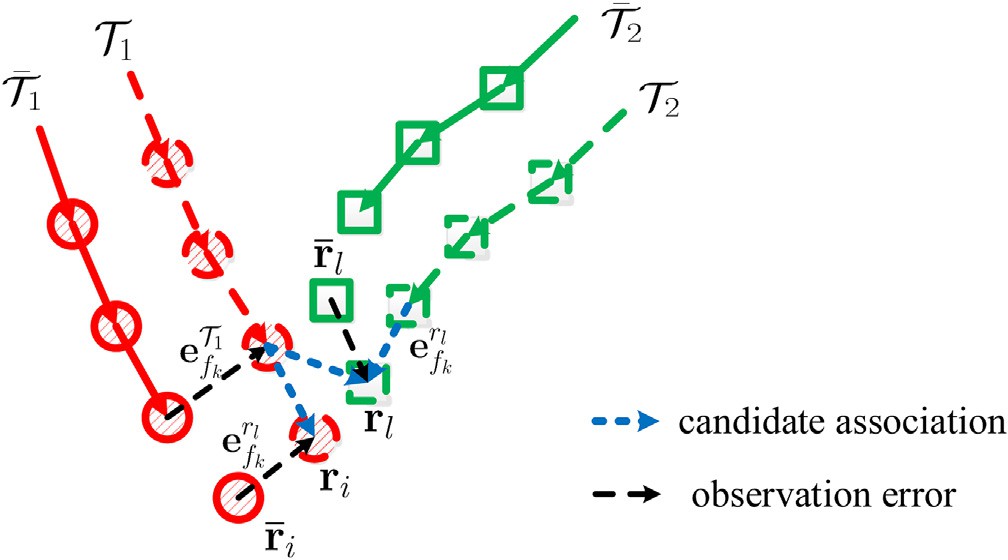
*fk* 27

Then the feature space 7 is reconstructed into *K* hierarchies obeying two rules:

* + - 1. Lower hierarchies of the feature space should be con- structed with features which demonstrate higher reliability on target representation. In other words, target represen- tation based on feature *fk* 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 7 *Hk* = 7 *Hk*—1 ∪{*fk*}.

A brief illustration of hierarchical feature space is given in

*Hk*

[Fig. 3](#_bookmark5), where {*ri*

*Hk*

} and {*Tj*

} denote detection responses and

tracklets to be associated in hierarchy *Hk* 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.

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

* + 1. *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](#_bookmark3). For *k*th hierarchy *Hk*, suppose there are *Mk* tracklets and *Nk* detection responses to be associated. Let У *H* := {У *j* } denote *Mk* tracklets and ℛ*H* := {*ri* }*N* denote *Nk* responses.

*Hk*

*Hk*

*k*

*k*

*k Mk*

First, an affinity matrix M *Hk* between У *Hk* and ℛ*Hk* is

calculated. Let A*ij* denote the element in the *i*th row and *j*th

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 У *j* in miss detection set *MHk* , causes about miss are analyzed. If miss detection is due to exit, У *j* is removed from У . If due to occlusion, an oc- clusion handling strategy proposed in our previous work [[47]](#_bookmark51)

column of M

*Hk ij*

is the affinity between *r*

and У

is adopted. It can effectively find reappearing response and use

*Hk* . A*Hk* *i*

considering all features {*fk*} in 7 *Hk* , given as:

*j*

it to update the tracklet У *j*. Conflicting links in set *CHk* are

A*ij* = *P*

*Hk*

*r* ; У 7

= Y

A *r* ; У (4)

transferred to the higher hierarchy *H*

*k*+1

to be further distin-

*fk* 27 *Hk*

where A *r* ; У = *G* Ð *r* (*t*); У ; m ; S (5)

*link*

*i*

*j*

*Hk*

*fk*

*i*

*j*

guished by combining more features in feature space 7 *Hk*+1 .

Finally, this iterative process is terminated until the last hier-

*fk*

*i*

*j*

*fk*

*i*

*j*

*fk*

*fk*

archy *HK* is processed or all conflicting links are distinguished.

For example in [Fig. 3](#_bookmark5), all conflicting links become reliable

Then, based on the affinity matrix M *Hk* , a hierarchical data association algorithm is conducted to handle data association on the *k*th hierarchy. The algorithm is given in the pseudo- code procedure in Algorithm 1. Its function is to find reli- able links *RHk* , conflicting links *CHk* , miss detections *MHk* and noise detections *NHk* in each hierarchy *Hk*. A brief illustration of the hierarchical data association is shown in [Fig. 3](#_bookmark5).

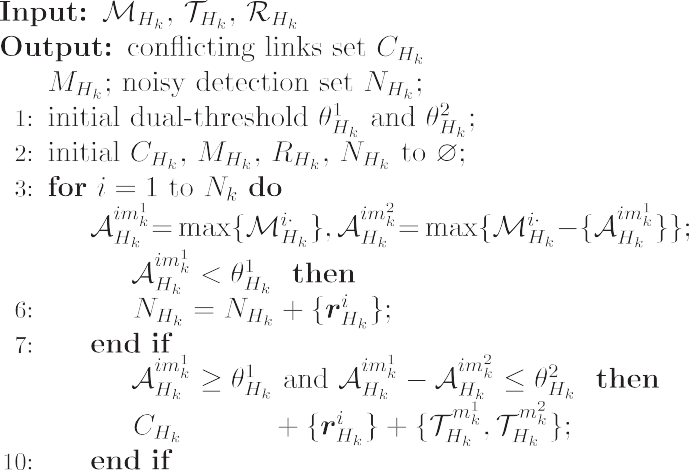
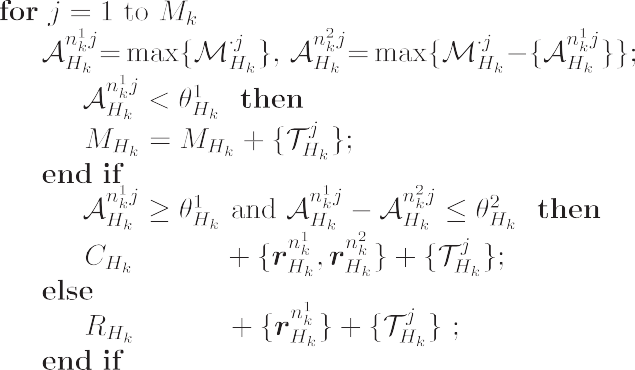
After that, reliable links in *RHk* are associated. For any noise detection in set *NHk* , a new tracklet is informally initialized

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 accumu- lation compared with the classic schemes. On the other hand several practical multi-tracking problems such as miss detec- tion, noise detection can be handled in this unified framework.











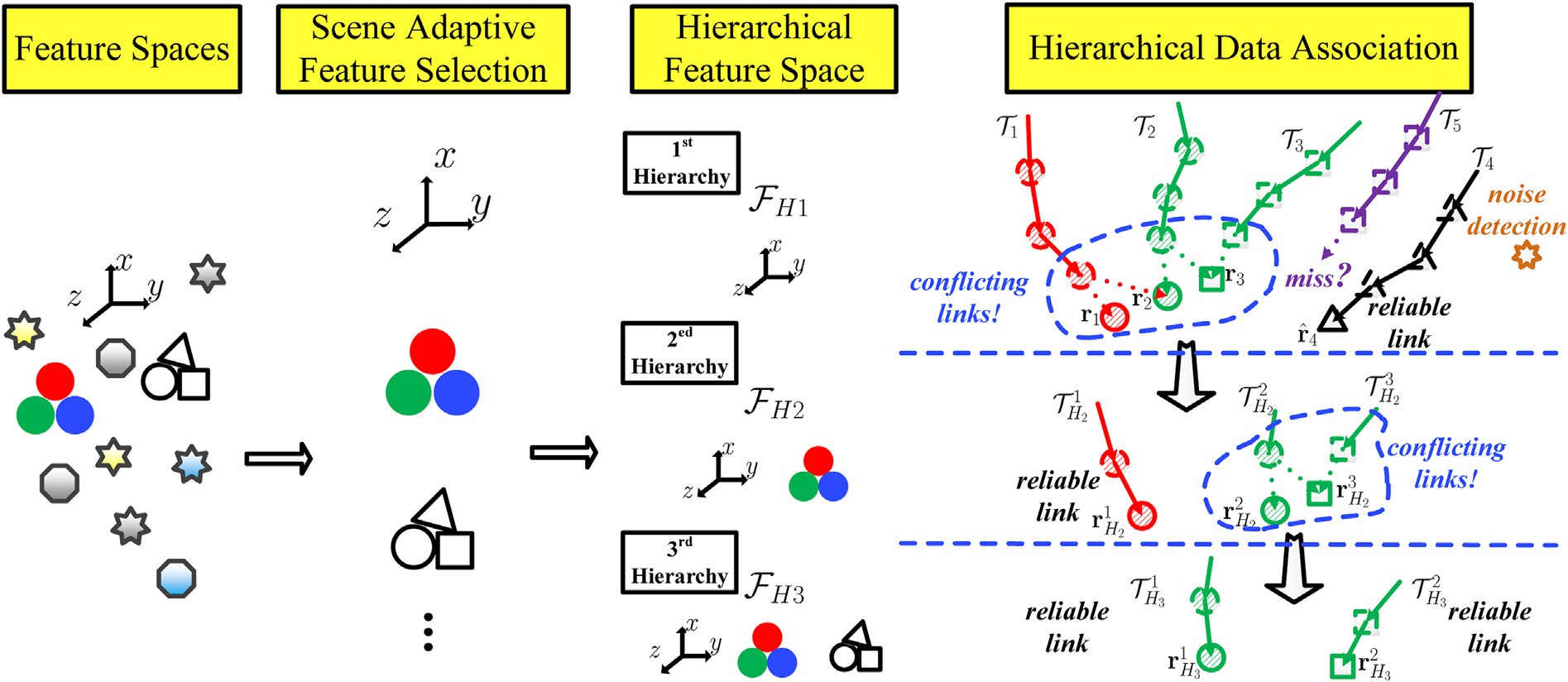


Fig. 3. A brief illustration of hierarchical data association on scene-adaptive feature space. Noise detection, miss detection, reliable links and conflicting links are unified and properly handled in this hierarchical framework.

* + 1. *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]](#_bookmark31), 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 *p*, motion state *m*, color *c*, appearance model *a* etc., and their general formulations are elaborated as follows:

Suppose for each tracklet У *j*, its latest response *rt*—1 can be denoted as:

*j*

As mentioned in the last section, given any detection response *ri* and tracklet У *j*, a user-defined metric *Ffk* ($) is used

for the calculation of affinity A*fk* (*ri*; У *j*) based on each feature

*fk*. Here, similar to previous related works such as [[14]](#_bookmark31), position affinity is assumed to obey a Gaussian distribution, which is also validated by observation in practice. So position affinity be- tween У *j* and a new detection response *ri* is calculated by:

A*p* *ri*; У *j* = *G* *pi* — *pt*—1 ; m*p*; S*p* (8)

*j*

where m*p* and S*p* are scene-adaptive parameters according to reliability of features in each scene, which will elaborated in detail in Section [3.3](#_bookmark8).

Similarly, details of motion cue and its affinity distance metric is given as follows. Motion *mj* of *tt*—1 in a tracklet У *j* is formulated using First-order Markov Model as:

*j*

*t*—1 *t*—2

*t*—1

n *t*—1

*t*—1 o

*mj* = *pj* — *pj* .

Motion affinity between У *j* and *ri* is also given obeying

Gaussian distribution with its own scene-adaptive parameters

where *pt*—1, *mt*—1, *cj* and *aj* indicate features of position, mo- m*m* and S*m*, formulated as:

*rj* =

*pj* ; *mj* ; *cj*; *aj*; /

(6)

*j j*

tion, 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 *pt*—1 is represented as Euclidean location

*t*—1 *j*

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

A*m* *ri*; У *j* = *G* ¨*pi* — *p* — *mj*¨; m*m*; S*m* (9)

*j j j* in multi-tracking.

(*Xc*; *Zc*) in camera coordinates based on depth data, in order to

avoid perspective effect in 2D image plane, formulated as:

Color feature *cj*

can be designed as histogram vectors of

*t*—1

*p*

*j*

*j j* *j*

*c*

*j Zj*

*c*

some commonly used color space such as HSV, RGB, YUV

etc., or combinations of them. Similarly, the metric of color

(7)

affinity between У *j* and *ri* is given by follows:

where *dj* is average depth of the detection response *rt*—1 on

*j*

=

*Xc*; *Z*

; *Zc*

= a*d*$*dj*; *X*

=

*c*

*fu*

*uj* — *u*0

1 !

depth image and a

*d*

*u*0 is the intrinsic parameters of the camera and *uj* is the width

is scale factor for quantization. Here *fu* and

A*c ri*; У *j* = *G*

corr *c* ; *c* ; m*c*; S*c*

(10)

coordinate on 2D image plane.

*i*

*j*

where *corr*(*vi*; *vj*) calculates correlation of vector *vi* and *vj*.

For appearance model, it is also regarded as a kind of high

representation is stable between *r*(*t*) and *r*(*t*—1), which is vital

*j j*

level features in the feature space. Due to its sophisticated

design, it will be elaborated in details in Section [4](#_bookmark11).

Besides these features mentioned above, other widely used

for data association. The variation between them is defined as:

*vj* (*t*)=Ð *r*(*t*); У (11)

*fk*

*fk*

*j*

*j*

features such as textures, edges, Haar-like features, interest

points, image patches, segmented regions [[44]](#_bookmark48) and also the combination of these basic features [[45]](#_bookmark49) [[46]](#_bookmark50) are also included in our feature pool.

* 1. *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 se- lection 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](#_bookmark4), 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:

Similarly, here Ð*fk* ($) represents distance metric between two detection responses under representation of feature *fk*.

A brief illustration of target representation variation is given in [Fig. 4](#_bookmark9), where three different features are selected for describing a same target in two scenes. It can be observed from [Fig. 4](#_bookmark9) 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 pro- posed and described as follows:

* + 1. *Reliability measurement*

In order to explore the feature's reliability for target rep- resentation in a given scene, two statistics of variation *vi* (*t*)

*f*

*k*

are studied:

One statistic is the mean of variation m*fk* , formulated as:

P*t* P*Nr* (*l*) *vj* (*l*)

P

m

= *l*=1 *j*=1 *fk*

(12)

*fk*

*t*

*l*=1

*Nr*(*l*)+ 1

First, for the convenience of the following description, let *ri*(*t*) denote detection response *ri* at frame *t*, and tracklet У *j* represents the set of all detection responses associated to target

*j* before frame *t*, written as У *j* := {/; *r*(*t*—2); *r*(*t*—1)}. So if *ri* is associated to at frame *t*, then *r* is *j* can *j* written as *r*(*t*).

У *j i* also be *j*

If *ri* has already been associated to У *j* at frame *t*, given

feature *fk*, we are interested in whether the feature

where *Nr*(*t*) is the number of detection responses associated to tracklets at time *t*. This statistic m*fk* essentially indicates the average true variation of feature *fk* on representing targets in a given scene. Lower true variation means higher stable repre- sentation for moving targets.

The other statistic of variation *vi* (*t*) is standard deviation of

*f*

*k*

variation *sfk* , formulated as:

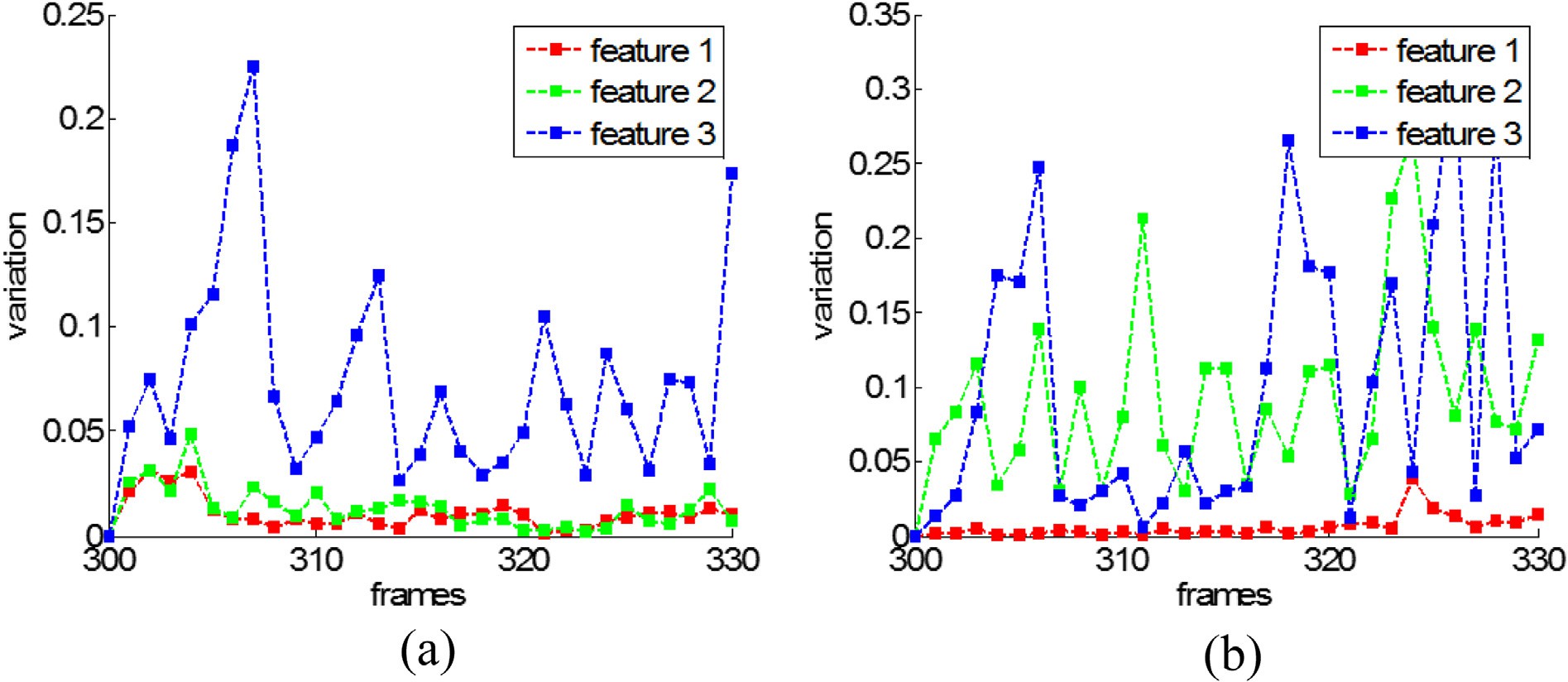


Fig. 4. A brief illustration of feature representation variations in two different scenes, where feature 1e3 are color histogram vectors in HSV (red line), RGB (green line) and YUV (blue line) color space respectively. It can be observed that feature 1 has lower variation in both scenes. Feature 2 has better performance in Scene 1 than Scene 2. Feature 3 has larger variation in both scenes.

1 here the parameter *Uf*

1

0

2

is also used to transform the hetero-

BP*t*

P*Nr* (*l*) *vj* (*l*) 2 — m

2C

*k*

geneous statistics of different features into homogeneous data

*fk* B@

*s* =

*l*=1

*j*=1 *t*

*l*=1

*fk fk*

*Nr*(*l*)+ 1 CA

P

(13)

to make sure they can be compared together. As mentioned above, m*fk* describes the average true variation of feature *fk* on representing same targets in consecutive frames, and *dfk* de- scribes the average distance between different targets in the

where *Nr*(*t*) is number of detection responses associated to

tracklets at time *t*. This statistic *sfk* essentially indicates sta- bility of feature feature *fk* on representing targets in a given scene. Higher observation errors brings higher value of *sfk* 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

*vj* (*t*) low values of mean m*f* and standard deviation *sf* .

given scene. Naturally, feature with relatively higher *dfk* but relatively lower m*fk* 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 m*pos* is relatively closed to the average position distance *dpos* compared with less crowded scenes. This indicates that position feature have more discriminative power in less

*fk k*

*k* crowded scenes than crowded scenes. Therefore, the dis-

In this framework, the reliability of feature *fk* on target

representation can be given as:

criminability measurement *Dk* essentially gives a criterion to select features to reduce ambiguity in data association for the

*Rk* = *Ufk*

1 + u $ 1

*ufk fk*

*r s*

(14)

scene.

Finally, in practice the reliability measurement *Rfk* and the discriminability measurement *Dfk* are combined to calculate

where *Ufk*

is a prior parameter for each feature *fk* to transform

the quality of feature *fk* for targets representation in the given

scene, formulated as follows:

the heterogeneous statistics of different features into homo-

geneous data to make sure reliability of different features can be compared together. Parameter u*r* is weighting factor to

*Qfk*

= lg *Rfk*

+ u*q*lg *Dfk*

(18)

control the impact of m*fk* and *sfk* for the reliability *Rfk* .

* + 1. *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 rela- tively too small, so the position feature is not a satisfactory feature. Suppose the distance between any two detection re- sponses under representation of feature *fk* at time *t* is given as:

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

1. Depth-invariant part-based appearance model

Besides some low level features mentioned in Section [3.2.3](#_bookmark6), 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]](#_bookmark39), pedestrian detection [[40](#_bookmark47)e[43]](#_bookmark47), multi- target tracking [[4,29,28]](#_bookmark25), 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

*ij*  (*t*) (*t*)

*vfk* (*t*)= Ð*fk*

*ri* , *rj*

discrimination. In this paper, a depth-invariant part-based

(15)

appearance model is proposed for target representation, which

Based on a certain period of observation for a given scene, another statistic *dfk* is introduced which is formulated as follows:

P*t* P*Nr* (*l*) P*Nr* (*l*) *vij* (*l*)

*d*

= *l*=1 *i*=1 *j*=*i*+1 *fk*

(16)

is described as follows:

* 1. *Depth invariant transform*

First, a depth-invariant transform (DIT) is applied to the

original image patch of the detection response, and transform

*fk* P*t*

*l*=1

*r*

*r*

*N* (*l*)(*N* (*l*)— 1)/2

where *Nr*(*t*) is number of detection responses associated to tracklets at time *t*. Intrinsically, the statistic *dfk* describes the

it to a novel depth-invariant image coordinates (DIIC). Sup-

pose the original image coordinates of the detection response is written as (*u*,*v*), and the depth-invariant image coordinates is written as (*u*, *v*), then the depth-invariant transformation can

average distance between detection responses in the given

scene, under the representation of feature *fk*.

Then, the discriminability measurement *Dfk* is proposed

taking two statistics *dfk* and m*fk* into consideration, formulated

as:

be formulatebd bas:

*u* = *a* $ *W* + *d*$(*u* — *u* )$1

*s*

*o*

0

*fu*

(19)

b

1

*D* =

*fk U*

*fk*

*dfk* — m*fk*

(17)

*v* = *a* $

1

*Ho* + *d*$(*v* — *v*0)$ $cos a*p* + *d*$sin a*p*

b *s*

*f*

*v*

where (*u*0, *v*0, *fu*, *fv*) is intrinsic parameters of RGB sensor, *as* is the scale factor, and *d* is the corresponding depth value of coordinate (*u*,*v*) on depth image.

Here, *Wo* is an offset to make sure that after DIT trans-

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.

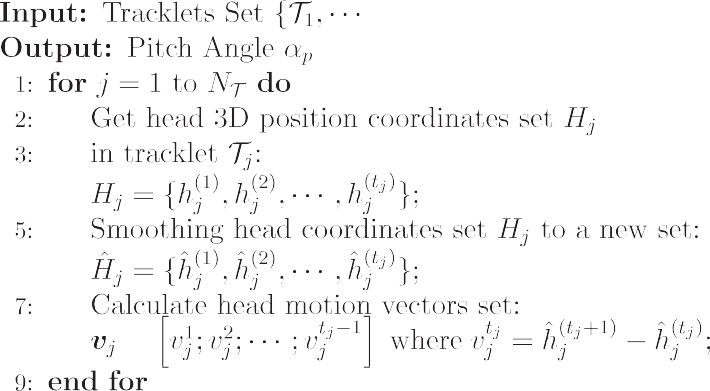
formation, and the minimum value of *u* is 1. And *H*

is set to

Here, we conduct a statistics of detection responses of each

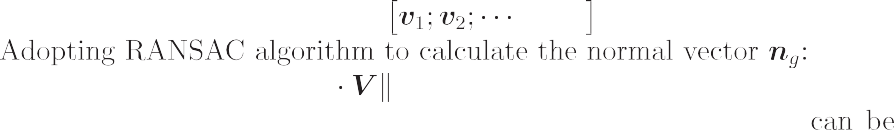
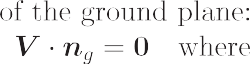
b *o*

make sure the top of all depth invariant patches correspond to a same height in the world coordinates. The parameter a*p* in [Formula (19)](#_bookmark12) is the pitch angle of the sensor. It can be seen from [Fig. 5](#_bookmark13) 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](#_bookmark15). Two detection re- sponses 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.





It is worth to mention that the parameter a*p* in [Formula (19)](#_bookmark12) 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]](#_bookmark55) usually calculate the normal vector of the ground plane and use it to estimate camera pitch angle. For example, work in [[51]](#_bookmark55) manually selects a portion of depth image area that corresponds to the ground plane. Although work in [[52]](#_bookmark56) 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



target based on a period of observation to the applied scene. A brief illustration of camera pitch angle estimation is given in [Fig. 6](#_bookmark14). 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 reli- able 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

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.

* 1. *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]](#_bookmark57). After the DIT transform, based on the detection response in the novel depth- invariant image coordinates, we combine basic features

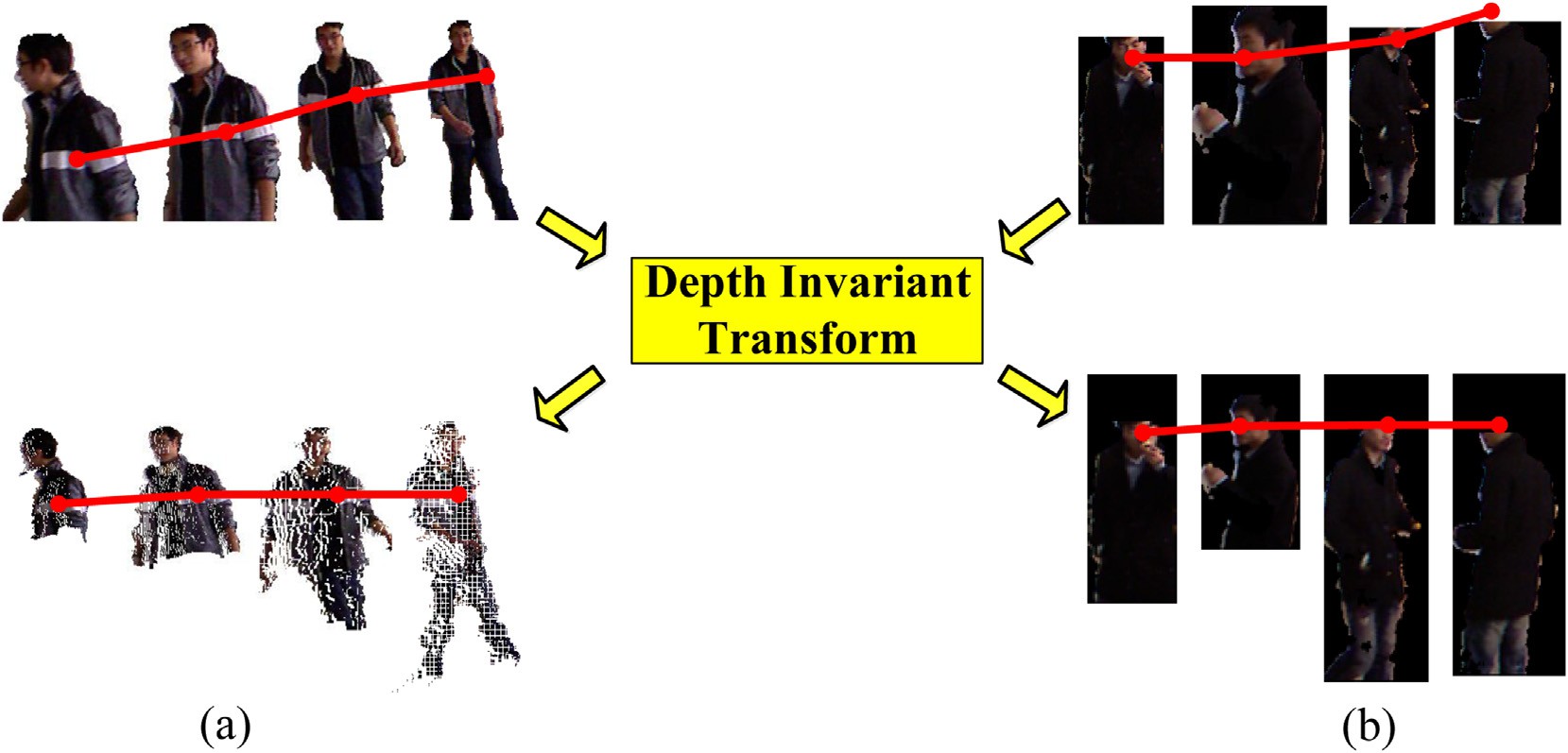


Fig. 5. The figure shows patches of the two targets' detection responses before and after DIT during the tracking process. Two versions of image patches with black and white backgrounds are shown to give a better visualization of the transformed patches. The red lines indicate same body parts in the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

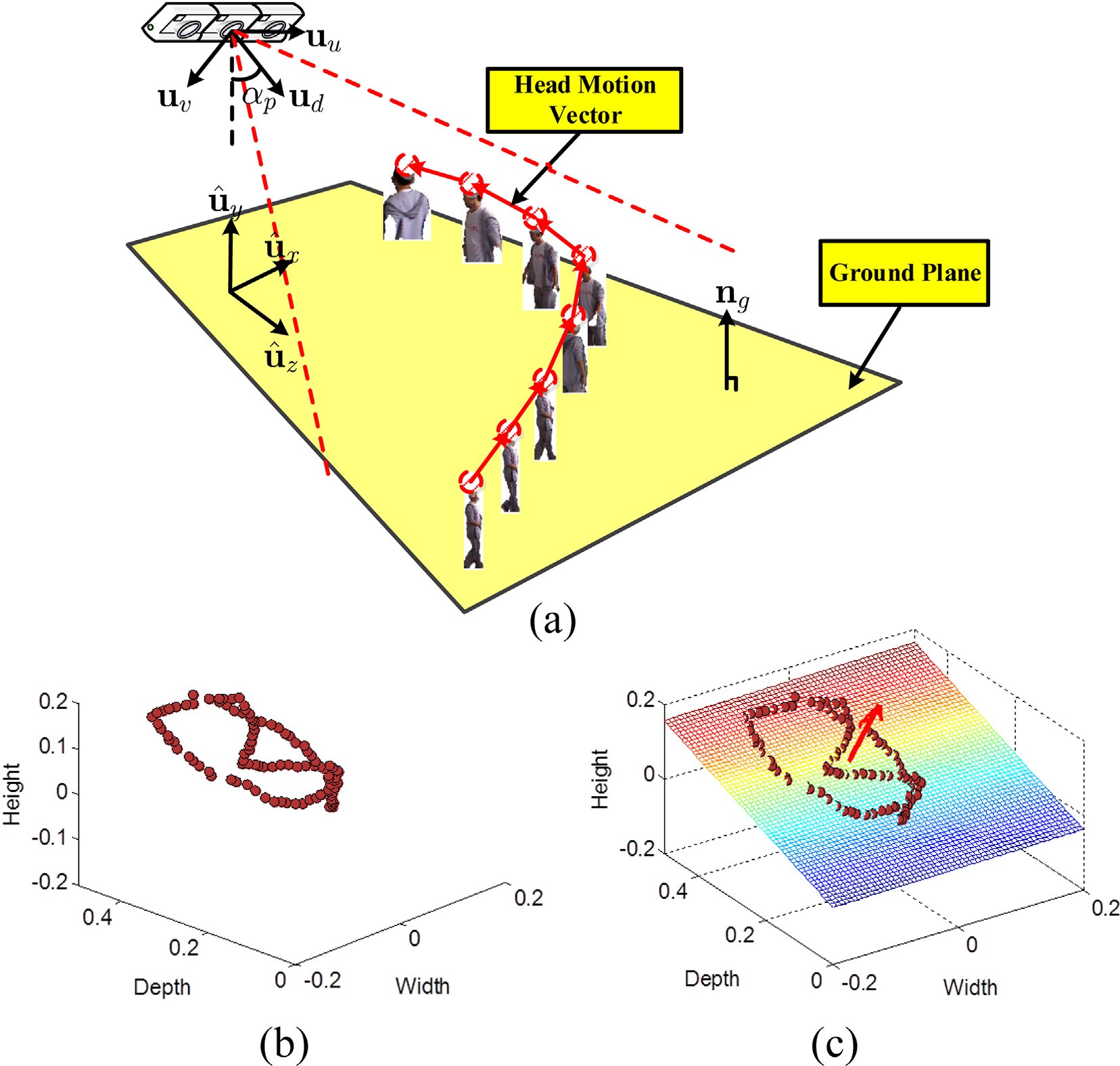


Fig. 6. A brief illustration of camera pitch angle estimation through long-term observation in the applied scene.

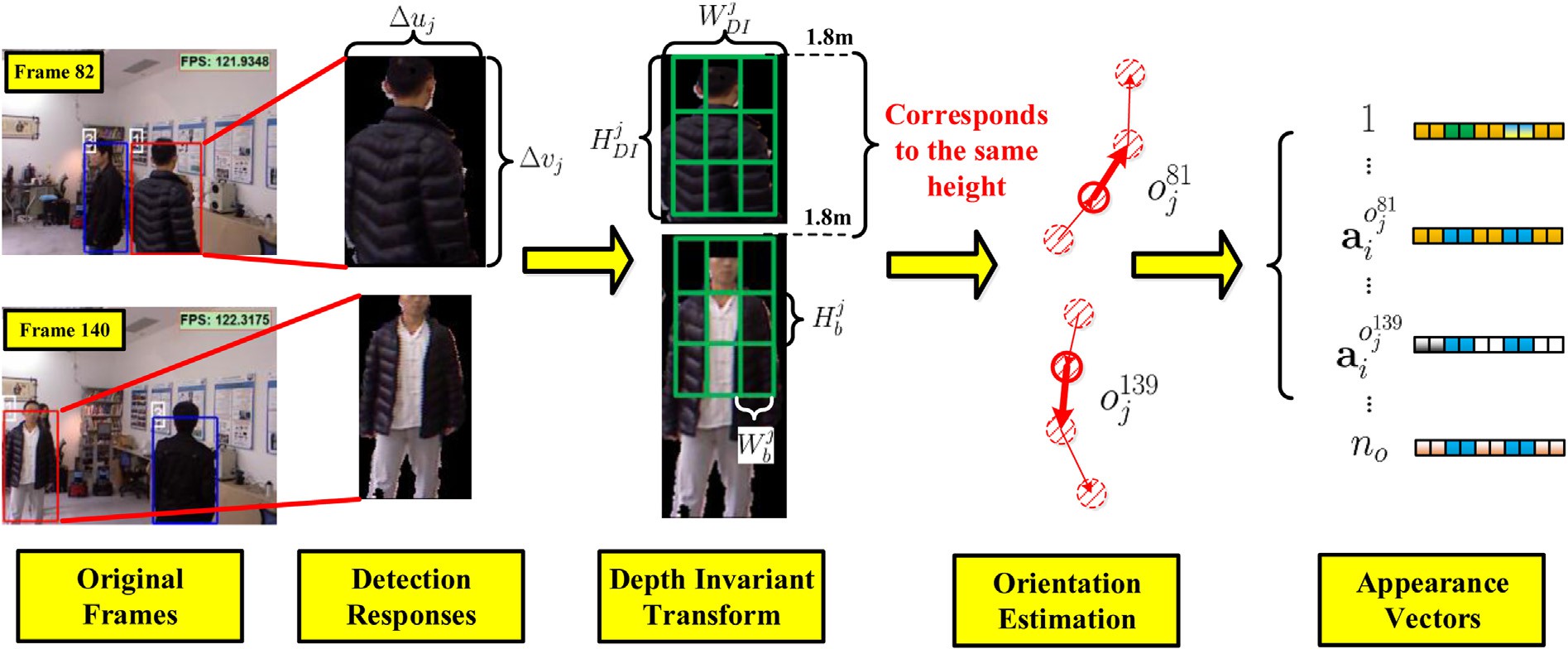


Fig. 7. Scale of the red target's detection responses changes in 2D image plane when target moves near and far, but it maintains relatively stable scale after Depth Invariant Transform (DIT).

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]](#_bookmark57), 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 infor- mation as suggested in [[54]](#_bookmark58), SIFT descriptors are used to capture texture information and handle illumination variation and LAB color histograms are extracted to enhance illumi- nation invariance.

Different from pervious part-based approaches, which

usually divide the image patch of detection response into

Variances m*fk* and S*fk* are given based on statistics mean m*fk* and standard division *sfk* during long term observation following the algorithm proposed in Section [3.3](#_bookmark8).

* 1. *Orientation-based modeling*

In indoor environments, appearance of a same target may vary a lot due to turning around, as shown in [Fig. 7](#_bookmark15), 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 У *j*, the coarse orientation of its response *tt*—1 at time *t*—1 is given by:

*j*

*Zct*—1 — *Zct*—2 !

*ot*—1 = s

arctan

*j*

*j*

2[1/*n* ] (21)

*j*

*no*

*o*

several part blocks in 2D image plane, In this work the block *j j*

*Xct*—1 — *Xct*—2

size of each part has a depth-invariant size (*Wj* , *Hj* ), which

*b b* where s*n* is a quantization function which quantify the

correspond to fixed values in the world coordinates, as shown

in [Fig. 7](#_bookmark15), 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 re- quirements of the applied scene. Assume the appearance

A

*r*

У

orientation of a response into *no* levels. Hence the appearance *aj* and color model *cj* of a tracklet У *j* are divided into *no* models {*a*1, /, *ano* } and {*c*1, /, *cno* } .

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

*j j j j*

1 !

= *G*

; m , S

(22)

corr

*c i* , *c j*

*bi* *i*

vector of a part block is denoted as *a*

, where *b* is the index of

*c i j*

*o ot*—1 *a c*

*j*

the blocks, the part-based model of each detection response *r*

can be further formulated as a concatenated vector

*aj* = (*ab*1, *ab*2, /, *abn*), and each *abn* is the appearance vector

*i j*

1 !

corr

*a i* , *a j*

of a part block and *bn* is number of part blocks.

and *a* ,

Thus, given two constructed appearance vector *a*

*i*

A*a ri* У *j* = *G*

*j*

*o ot*—1

*i*

*j*

; m*a*, S*a*

(23)

appearance affinity between У *j* and *ri* can also be given as the following general form:

1 !

where *corr*(*vi*, *vj*) calculates the correlation between vectors *vi*

and *vj*.

A*a ri* У *j* = *G*

*i*

*j*

corr *a* , *a* ; m*a*, S*a*

(20)

In practice application, it is also a scene-adaptive problem

In this framework, similar to the related work in [[14]](#_bookmark31), features' affinities are all defined by Gaussian distributions.

whether adopting multiple or single orientation appearance

model. For example, in some scenes targets always move to- wards 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 re- sponses always have multiple orientations.

Table 1

Detailed configuration of experimental datasets. Dataset name Dataset type

In/Out Distance Crowdedness RGB-D/RGB

Scene 1 e Scene 5 Indoors Close Crowded RGB-D Scene 6 e Scene 10 Indoors Close Crowded RGB Scene 11 e Scene 15 Outdoors Distant Less Crowded RGB

1. Experiment and analysis

PETS09 S2.L1 e S2.L3 [[49]](#_bookmark53)

Outdoors Distant Less Crowded RGB

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](#_bookmark17), 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 frame- work 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 perfor- mance of the depth invariant part-based appearance model (DIAM) in various scenes, which is given in Section [5.3](#_bookmark20). Third, we analyze advantages and disadvantages of the pro- posed 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](#_bookmark22).

* 1. *Datasets and settings*

TUD campus [[50]](#_bookmark54) Outdoors Close Crowded RGB TUD crossing [[50]](#_bookmark54) Outdoors Close Crowded RGB

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 associ- ated sequences, two statistics *ufk* , *sfk* and reliability *Rfk* for all features in feature space are calculated according to algorithm

given in Section [3.3](#_bookmark8). Due to its statistical nature, occasional id- switches can be overlooked in data association. The variation reference *Ufk* in [Formula (14)](#_bookmark10) can be obtained by averaging variations of each feature *fk* during association in several se- quences in different scenes. It is a constant value based on prior statistics which is scene-independent. The weighting factor u*r* is empirically set to 0.5 to control the impact of m*fk*

and *sfk* for the reliability *Rfk* in [Formula (14)](#_bookmark10).

For our approach, a and *as* is set to 10 in [Formulas (7) and](#_bookmark7)

[(19)](#_bookmark7). *Hoff* is set to 100 in [Formula (19)](#_bookmark12). For part-based appearance model, parts number *bn* is set to 9 and block

size {*Hj* , *Wj* } (shown in [Fig. 7](#_bookmark15)) is set to {30, 20}, due to

*b b*

There is no generally accepted benchmark available for multi-tracking [[5]](#_bookmark26). Therefore, we recorded our own multi- tracking dataset including both RGB and RGB-D data, in- door 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

depth-invariant nature, these parameters correspond to abso-

lute values in the world coordinates, they are relatively uni- versal in the human tracking system (see [Fig. 8](#_bookmark18)).

* 1. *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]](#_bookmark52), defined as

P*t* *cm*(*mt*)+ *cf* (*fpt* + *cs*(*mmet*))

includes many 3D features. All these modules need both RGB and depth data. For the RGB datasets, we only use it to

*MOTA* = 1 —

P*tgt*

(24)

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 e S2.L3 [[49]](#_bookmark53), TUD Campus [[50]](#_bookmark54), TUD Crossing [[50]](#_bookmark54). The detailed configuration of the experimental datasets is given in [Table 1](#_bookmark16).

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

where *gt* is the number of ground truth detections, *mt* the number of miss-detections, *fpt* is the false positive count and *mmet* the number of instantaneous identity switches [[12]](#_bookmark30). 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](#_bookmark23). Here non- HDA stands for combining all features in feature space for data association. It can be seen from [Table 2](#_bookmark19) 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

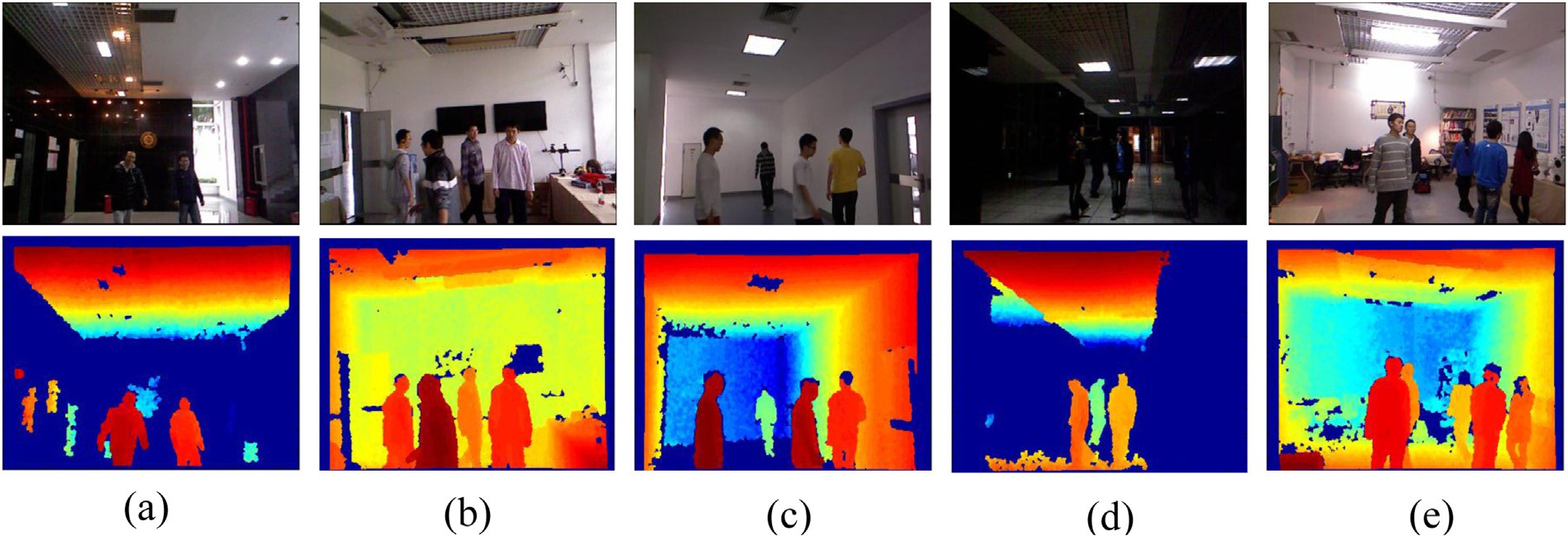


Fig. 8. Figures from (a) to (e) are samples of our RGB-D datasets recorded in Scene 1 to Scene 5.

Table 2

Quantitative accuracy and speed comparison (not including detection time, only data association) of several data association schemes in our datasets ([Table 1](#_bookmark16)).

Frameworks MOTA (%)/FPS

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Scene 1e5 | Scene 6e10 | Scene 11e15 | 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 compu- tation 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](#_bookmark19) 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.

* 1. *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 experi- ments are conducted with classic average-division part-based appearance model (ADPAM) [[28]](#_bookmark38), and a state-of-the-art classifier-based discriminative appearance model (CDAM) [[4]](#_bookmark25), 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]](#_bookmark38), and the other is a state-of-the-art classifier-based discriminative

appearance model (CDAM) proposed in [[4]](#_bookmark25) which combines color histograms, covariance matrixes, and histogram of gra- dients (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](#_bookmark21). 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 vulner- able to large size variation and truncation by the field-of-view. Take detection responses shown in [Fig. 5](#_bookmark13) 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]](#_bookmark38) | 76.5/20.2 | 70.2/22.5 | 65.2/23.0 | 78.2/21.5 | 61.5/19.5 |
| CDAM [[4]](#_bookmark25) | 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](#_bookmark17), 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 hierar- chical 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](#_bookmark23), various classic features for target representation are used for feature space construc- tion, including position (P), size (S), motion (M), color (C) and appearance model (A), which have been elaborated in Section [3.2.3](#_bookmark6) and Section [4](#_bookmark11). 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](#_bookmark23).

It can be seen from [Table 4](#_bookmark23) that different feature permu- tations 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 illumi- nation 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

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.

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 applica- tions in various scenes.

Then we evaluate advantages of the hierarchical data as- sociation scheme (HDA) to the non-hierarchical data associ- ation scheme (non-HDA). Quantitative results in form of MOTA (%) and FPS are given in the bottom part of the [Table](#_bookmark23)

[4](#_bookmark23). Compared with the HDA results in [Table 4](#_bookmark23), 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 associa- tion. 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.

1. 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 grad- ually fuses more features according to requirements of dis- tinguishing 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 con- tributes 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)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref1)

[(Jan. 2011) 120](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref1)e[129](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref1).

1. [P. Rashidi, D.J. Cook, L.B. Holder, S.E. Maureen, IEEE Trans. Knowl.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref2) [Data Eng. TKDE 23 (4) (Apr. 2011) 527](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref2)e[539](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref2).
2. [J. Han, E.J. Pauwels, P.M. de Zeeuw, P.H.N. de With, IEEE Trans.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref3) [Consumer Electron. 58 (2) (May 2012) 253](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref3)e[263](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref3).
3. [C. Kuo, C. Huang, R. Nevatia, IEEE Conf. Comput. Vis. Pattern Rec-](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref4) [ognit. CVPR (2010) 685](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref4)e[692](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref4).
4. [M.D. Breitenstein, F. Reichlin, B. Leibe, E.K. Meier, L.V. Gool, IEEE](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref5) [Trans. Pattern Anal. Mach. Intell. TPAMI 33 (9) (Sept. 2005)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref5) [1820](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref5)e[1832](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref5).
5. [C.X. Liu, S.G. Gong, C.C. Loy, X.G. Lin, Eur. Conf. Comput. Vis. Work.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref6)

[ECCVW (2012) 391](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref6)e[401](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref6).

1. [B. Prosser, W. Zheng, S. Gong, T. Xiang, Br. Mach. Vis. Conf. BMVC 1](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref7)

[(3) (2010) 1](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref7)e[11](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref7).

1. [W.S. Zheng, S.G. Gong, T. Xiang, IEEE Conf. Comput. Vis. Pattern](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref8) [Recognit. CVPR (2011) 649](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref8)e[656](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref8).
2. [J. Berclaz, F. Fleuret, P. Fua, IEEE Conf. Comput. Vis. Pattern Recognit.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref9) [CVPR (2006) 744](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref9)e[750](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref9).
3. [J. Xing, H. Ai, S. Lao, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref10)

[(2009) pp.1200](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref10)e[1207](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref10).

1. [L. Zhang, Y. Li, R. Nevatia, IEEE Conf. Comput. Vis. Pattern Recognit.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref11) [CVPR (2008) 1](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref11)e[8](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref11).
2. [J. Berclaz, F. Fleuret, E. Turetken, P. Fua, IEEE Int. Conf. Comput. Vis.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref12) [ICCV (2011) 137](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref12)e[144](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref12).
3. [J. Berclaz, F. Fleuret, E. Turetken, P. Fua, IEEE Trans. Pattern Analy.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref13) [Mach. Intell. TPAMI 33 (9) (Sept. 2011) 1806](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref13)e[1819](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref13).
4. [C. Huang, B. Wu, R. Nevatia, Eur. Conf. Comput. Vis. ECCV (2008)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref14) [788](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref14)e[801](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref14).
5. [M.D. Breitenstein, F. Reichlin, B. Leibe, E. Koller-Meier, L.V. Gool,](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref15) [IEEE Int. Conf. Comput. Vis. ICCV (2009) 1515](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref15)e[1522](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref15).
6. [Y. Cai, N. de Freitas, J.J. Little, Eur. Conf. Comput. Vis. ECCV (2006)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref16) [107](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref16)e[118](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref16).
7. [K. Okuma, A. Taleghani, O.D. Freitas, J.J. Little, D.G. Lowe, Eur. Conf.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref17) [Comput. Vis. ECCV (2004) 28](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref17)e[39](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref17).
8. [B. Wu, R. Nevatia, Int. J. Comput. Vis. IJCV 75 (2) (Nov. 2007)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref18) [247](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref18)e[266](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref18).
9. [D.R. Magee, Image Vis. Comput. (2004) 143](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref19)e[155](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref19).
10. [Z. Khan, T. Balch, F. Dellaert, IEEE Trans. Pattern Anal. Mach. Intell.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref20) [TPAMI 27 (11) (Nov. 2005) 1805](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref20)e[1891](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref20).
11. [W. Choi, C. Pantofaru, S. Savarese, IEEE Int. Conf. Comput. Vis. Work.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref21) [ICCVW (2011) 1076](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref21)e[1083](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref21).
12. [M. Yang, F. Lv, W. Xu, Y. Gong, IEEE Int. Conf. Comput. Vis. ICCV](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref22) [(2009) 1554](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref22)e[1561](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref22).
13. [P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, IEEE Trans.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref23) [Pattern Anal. Mach. Intell. TPAMI 32 (9) (2010) 1627](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref23)e[1645](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref23).
14. [M. Hofmann, M. Haag, G. Rigoll, IEEE Int. Workshop Perform. Eval.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref24) [Track. Surveillance PETS (2013) 22](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref24)e[28](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref24).
15. [Hoan Nguyen, T. Fasciano, D. Charbonneau, A. Dornhaus, M.C. Shin,](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref25) [IEEE Winter Conf. Appl. Comput. Vis. WACV (2014) 941](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref25)e[946](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref25).
16. [M. Baum, P. Willett, IEEE Int. Conf. Acoust. Speech Signal Process.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref26) [ICASSP (2014) 4209](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref26)e[4213](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref26).
17. [B. Yang, R. Nevatia, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref27) [(2012) 2034](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref27)e[2041](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref27).
18. [B. Yang, R. Nevatia, Eur. Conf. Comput. Vis. ECCV (2012) 484](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref28)e[498](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref28).
19. [B. Yang, R. Nevatia, IEEE Conf. Comput. Vis. Pattern Recognit. CVPR](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref29) [(2012) 1918](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref29)e[1925](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref29).
20. [L. Bazzani, M. Cristani, V. Murino, Comput. Vis. Image Underst. CVIU](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref30) [117 (2) (2013) 130](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref30)e[144](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref30).
21. [N. Bird, O. Masoud, N. Papanikolopoulos, A. Isaacs, IEEE Trans.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref31) [Intelligent Transp. Syst. TITS 6 (2) (June 2005) 167](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref31)e[177](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref31).
22. [C. Wang, H. Liu, L. Ma, IEEE Signal Process. Lett. SPL 21 (6) (Jun.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref32) [2014) 717](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref32)e[721](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref32).
23. [B. Ni, N.C. Dat, P. Moulin, IEEE Int. Conf. Acoust. Speech Signal](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref33) [Process. ICASSP (2012) 1405](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref33)e[1408](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref33).
24. [J. Shotton, et al., Commun. ACM (2013) 116](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref34)e[124](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref34).
25. [L. Spinello, K.O. Arras, IEEE Int. Conf. Robotics Automation ICRA](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref35) [(2012) 4469](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref35)e[4474](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref35).
26. [L. Spinello, K.O. Arras, IEEE/RSJ Int. Conf. Intelligent Robots Syst.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref36) [IROS (2011) 3838](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref36)e[3843](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref36).
27. [L. Cruz, Patterns Images Tutorials SIBGRAPI (2012) pp.36](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref37)e[49](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref37).
28. [F. Sanchez, L. Rubio, J. Diaz, E. Ros, Mach. Vis. Appl. MVA 25 (5) (Oct.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref38)

[2013) 1211](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref38)e[1225](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref38).

1. [C. Liu, S. Gong, C.C. Loy, Pattern Recognit. 47 (4) (2014) 1602](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref39)e[1615](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref39).
2. [J. Marin, D. Vazquez, D. Geronimo, A.M. Lopez, IEEE Conf. Comput.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref40) [Vis. Pattern Recognit. CVPR (2010) 137](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref40)e[144](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref40).
3. [M. Liem, D.M. Gavrila, IEEE Int. Conf. Work. Automatic Face Gesture](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref41) [Recognit. (2013) 1](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref41)e[6](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref41).
4. [P. Sermanet, K. Kavukcuoglu, S. Chintala, Y. LeCun, IEEE Conf.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref42) [Comput. Vis. Pattern Recognit. CVPR (2013) 3626](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref42)e[3633](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref42).
5. [M. Tang, Andriluka, B. Schiele, Int. J. Comput. Vis. IJCV (2014)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref43).
6. [S. Bak, E. Corvee, F. Bremond, M. Thonnat, IEEE Int. Conf. Adv. Video](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref44) [Signal Based Surveillance AVSS (2010) 1](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref44)e[8](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref44).
7. [L. Bazzani, M. Cristani, A. Perina, V. Murino, Pattern Recognit. Lett.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref45) [PRL 33 (7) (May 2012) 898](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref45)e[903](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref45).
8. [Y. Lu, L. Lin, W.S. Zheng, IEEE Int. Conf. Comput. Vis. CVPR (2013)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref46) [3152](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref46)e[3159](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref46).
9. [H. Liu, Z. Yu, H.B. Zha, L. Zhang, Y.X. Zou, Pattern Recognit. Lett. PRL](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref47)

[30 (9) (July 2009) 827](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref47)e[837](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref47).

1. [K. Bernardin, R. Stiefelhagen, EURASIP J. Image Video Process. JIVP](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref48) [2008 (1) (Jan. 2008) 1](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref48)e[10](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref48).
2. [J. Ferryman, IEEE Workshop Perform. Eval. Track. Surveillance (2009)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref49).
3. [M. Andriluka, S. Roth, B. Schiele, IEEE Conf. Comput. Vis. Pattern](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref50) [Recognit. CVPR (2008) 1515](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref50)e[1522](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref50).
4. [A. Albiol, J. Oliver, J.M. Mossi, IET Comput. Vis. 6 (5) (Nov. 2012)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref51) [378](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref51)e[387](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref51).
5. [W. von Hansen, Photogramm. Image Anal. PIA (2007) 93](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref52)e[97](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref52).
6. [H. Liu, M. Qian, C. Wang, IEEE Int. Conf. Acoust. Speech Signal](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref53) [Process. ICASSP (2015)](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref53).
7. [M. Farenzena, L. Bazzani, A. Perina, V. Murino, M. Cristani, IEEE Conf.](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref54) [Comput. Vis. Pattern Recognit. (CVPR) (2010) 2360](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref54)e[2367](http://refhub.elsevier.com/S2468-2322(16)30065-8/sref54).

Prof. Hong Liu received the Ph.D. degree in me- chanical 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 sup- ported by “National High-level Talents Special Sup- port 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 pub- lished more than 150 papers and gained Chinese Na-

tional Aero-space Award, Wu Wenjun Award on Artificial Intelligence, Excellence Teaching Award, and Candidates of Top Ten Outstanding Pro- fessors 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 re- viewers 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 recon- struction, facial expression and gender recognition. He has published articles in IEEE International Confer- ence on Image Processing (ICIP).