RGB-D Based Multi-attribute People Search in Intelligent Visual Surveillance

Wu Liu[1,2], Tian Xia[1], Ji Wan[1,2], Yongdong Zhang[1], and Jintao Li[1]

[1] Institute of Computing Technology, Chinese Academy of Science, Beijing 100190, China
[2] Graduate School of the Chinese Academy of Science, Beijing 100039, China
{liuwu,txia,wanji,zhyd,jtli}@ict.ac.cn

Abstract. Searching people in surveillance videos is a typical task in intelligent visual surveillance (IVS). However, current IVS techniques can hardly handle multi-attribute queries, which is a natural way of finding people in real-world. The challenges arise from the extraction of multiple attributes which largely suffer from illumination change, shadow and complicated background in the real-world surveillance environments. In this paper, we investigate how these challenges can be addressed when IVS is equipped with RGB-D information obtained by an RGB-D camera. With the RGB-D information, we propose methods that accurately and robustly segment human region and extract three groups of attributes including biometrical attributes, appearance attributes and motion attributes. Furthermore, we introduce a novel IVS system which is capable of handling multi-attribute queries for searching people in surveillance videos. Experimental evaluations demonstrate the effectiveness of the proposed method and system, and also the promising applications of bringing RGB-D information into IVS.

Keywords: IVS, RGB-D, Multi-Attribute Query, People Search.

1 Introduction

Over the last decade we have witnessed an explosive growth of surveillance video data. This drives the birth of intelligent visual surveillance (IVS), which mainly aims at automatically detecting targets in surveillance video via computer vision techniques [1]. Searching people in a long surveillance video is a typical task in IVS, and a natural way of finding people in real-world is through a multi-attribute based query. For example, finding the person whose height is between 175cm and 180cm, with white skin, green T-shirt, blue shorts and a big black luggage, running across the hall. Unfortunately, existing IVS techniques can hardly handle the above multi-attribute query due to the difficulties in the extraction of multiple attributes. The bottleneck is mainly derived from environment modeling and object segmentation in IVS, which lay the foundation of IVS but largely suffer from illumination change, reflection, shadow and complicated background in the real-world surveillance environments [1].

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 750–760, 2012.

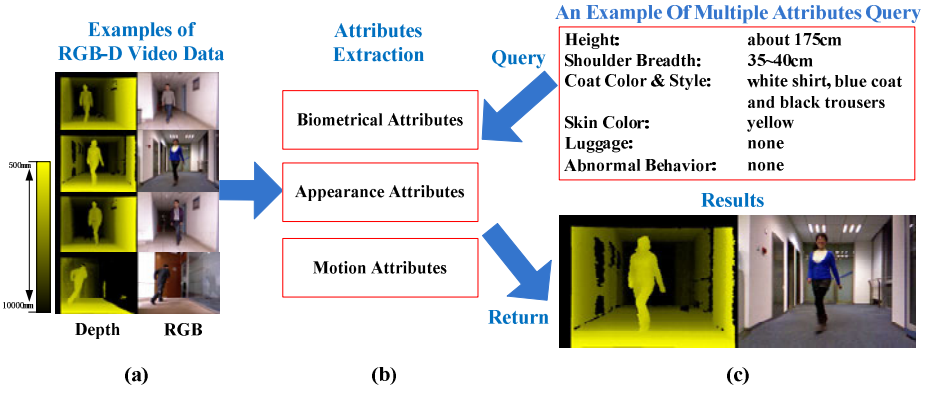

Fig. 1. An example of our IVS system which is capable of handling multi-attribute queries for searching people in surveillance video. (a) shows some example videos in our RGB-D database which is captured by a RGB-D camera. (b) lists the multiple attributes extracted from the RGB-D video data. (c) illustrates an example of multiple attributes query for people search in surveillance video and the retrieval result . (Best seen in color)

In this paper, we focus on the extraction of multiple attributes, and build an IVS system which is capable of handling the above multi-attribute query for people search in surveillance video. The improvement is largely derived from the usage of RGB-D information.

RGB-D is a key terminology in RGB-D project [2] whose goal is to develop techniques enabling future use cases of depth cameras (we name it RGB-D camera). RGB-D camera can capture RGB image along with its per-pixel depth information, as shown in Fig. 1. The additional depth information brings an opportunity to address a range of challenging issues. For example, one work of RGB-D project [3] is to use RGB-D information in robotic manipulation and interaction, and Microsoft Kinect [4] used it in human-computer interaction. RGB-D information also brings great opportunities to IVS, particularly for the challenging issues of object segmentation. The discontinuity of depth information can be utilized to perform nearly perfect object segmentation which avoids environment modeling and is invariant to illumination changes. Moreover, RGB-D information is helpful for us to get the 3D information of the scene, and makes it possible to extract more useful attributes which are provide powerful support to multi-attribute query for people search.

In order to investigate and verify the benefits of bringing RGB-D into IVS, we extract three groups of attributes, i.e., biometrical attributes including height and shoulder breadth; appearance attributes including skin color, clothing color and style, and luggage; motion attributes including squatting, running and wandering. These attributes are widely used in describing a person in real-world people finding. Based on these attributes, a novel IVS system with RGB-D camera is built to provide an effective way of finding people via multi-attribute query. Some searching results of multi-attribute queries from our IVS system are illustrated in Fig. 1.

The previous arts we know in searching people via visual attributes are [5] and [6]. [5] mainly focuses on facial and clothing color attributes in surveillance environment.

Although facial information is very useful in personal identification, its reliability is limited by the requirement of people's close shot, which is only feasible in some particular surveillance environment. [6] proposes an approach for ranking and retrieval of images based on multi-attribute queries. Since it is for static portrait data, its attributes are quite different from ours for surveillance video. [7] also proposes a multiple attributes matching method for video retrieval, without introducing the method of multi-attribute extraction.

Our contributions are summarized as follows:

(1) Taking the advantages of RGB-D information in human segmentation and 3D scene establishment, multiple attributes facilitating for finding people in surveillance environments are proposed. And their robust extraction is implemented effectively and efficiently, which demonstrates the validity of bringing RGB-D information into IVS.

(2) A novel IVS system with RGB-D camera is established, which provides an effective way of searching people via multi-attribute query, following the natural way of finding people in real-world.

2 System Overview

In this section, we present an overview of the proposed framework for people searching in surveillance video with RGB-D information. As shown in Fig. 2, the framework consists of four components: RGB-D camera, analytic engine, data storage and user search interface.

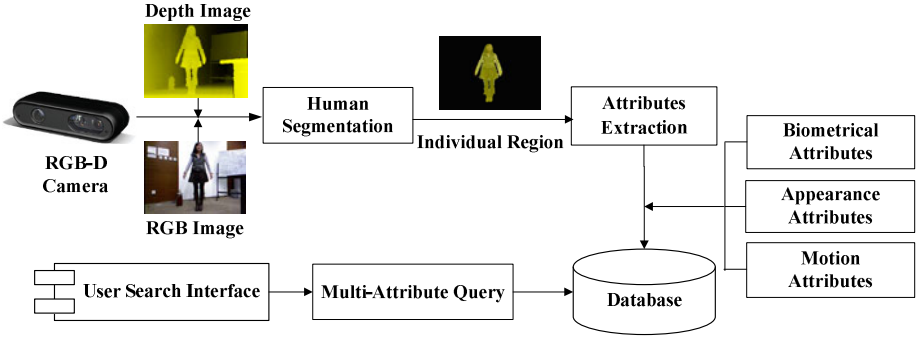

Fig. 2. The overview of our system architecture

2.1 RGB-D Camera

In our IVS system, the RGB-D camera uses Primesense [8] depth camera technology, it simultaneously captures both RGB and depth images at 640×480 resolution with 30 fps. Its field of view is 58° H, 45° V and 70° D for horizontal, vertical and diagonal perspective separately. Using OpenNI [9], we can calibrate the RGB and depth images.

2.2 Attributes Extraction

This module performs all the computer vision analysis in IVS system, and consists of three steps: (1) segmenting each individual region from the background; (2) extracting biometrical and appearance attributes for each individual including height, shoulder breadth, skin color, clothing color and style, luggage; (3) extracting motion attributes by detecting the abnormal behaviors including squatting, running and wandering. The details of this module are presented in Section 3.

2.3 Data Storage and User Search Interface

The extracted attributes are indexed into a relational database to facilitate efficient multi-attribute search. A novel user search interface is also provided, where users can input multi-attribute queries. As shown in Fig. 3, region A and B are for inputting multi-attribute queries. In region A, users can express their desired attributes concerning height, shoulder breadth and abnormal behaviors. In region B, users can express their desired attributes concerning skin color, luggage, clothing color and style by two operations, selecting color and filling in the human template. Region C shows the search results and region D is a playback window for surveillance videos.

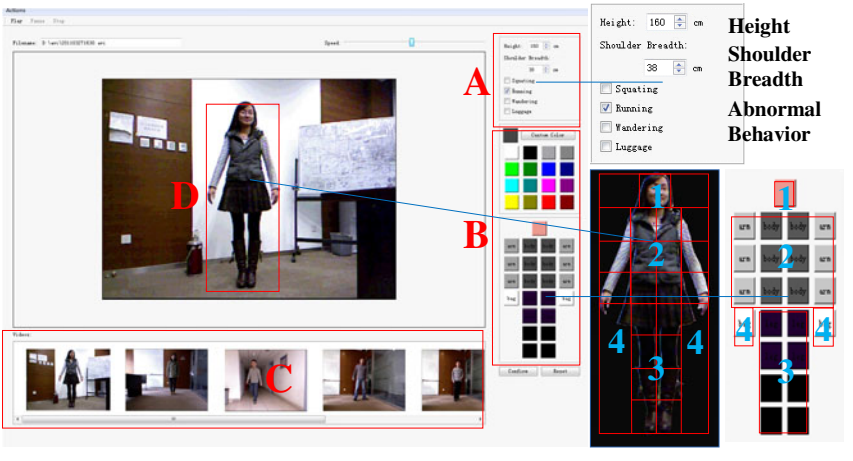

Fig. 3. The user search interface of our system. Region A and B are for inputting multi-attribute queries. Region C shows the search results and region D is a playback window for surveillance video. (Best seen in color)

3 Attributes Extraction

3.1 Human Segmentation

Our human segmentation is to take advantages of the RGB and depth information to obtain individual region as shown in Fig. 2. First, connected motion regions are

obtained based on RGB-D information; then, the refined individual region is identified from these regions. The details are shown as below.

Supposing there are n persons in frame I and we define $I = B \bigcup \{P_i\}_{i=1}^{n} \bigcup NP$ where B is the background, P_i is the individual region of human i, and NP is the remaining region apart from background and human, such as some moving objects.

Let $V_{(x,y)}(c,d)$ denotes the RGB-D information of the pixel at (x, y) in frame I, c denotes the RGB values and d denotes the depth value. We can simply remove the majority of background by setting the threshold $d_{\min}$ and $d_{\max}$ according to real monitoring conditions and only considering the region within $d_{\min}$ and $d_{\max}$. Then we use the Split-and-Merge strategy to slice the foreground into depth connected regions R_i. Extended temporal differencing method considering color and depth changes is utilized to obtain the connected motion regions R_i'.

As $\{P_i\}_{i=1}^{n} \bigcup NP \in R_i'$, in order to remove *NP*, the actual area measure S_r of each R_i' in real world is calculated by Eq. (1)

$$S_r = (S_p / S) \times 4d^2 \times \tan(\frac{hor}{2}) \times \tan(\frac{ver}{2}) \quad , \tag{1}$$

where S_p is the number of pixels in region R_i' and S is the total number of pixels in depth image. d is the average depth of region R_i', hor and ver are the RGB-D camera's horizontal and vertical field of view. We suppose the horizontal bisecting line of the camera image is horizontal in the real world as well. Then we can remove some regions which are too larger or smaller than real human area with S_r value.

In order to remove the remaining NP whose actual area is similar with real human, a cascade of Adaboost classifier based on Haar-like features is used to detect the head and a Bayes color model [11] is used to detect skin in the region R_i'. The detected regions are considered as seed location to refine the whole body contours from the region R_i'.

In order to demonstrate that RGB-D information can enhance the performance of human segmentation, we conduct comparison to background subtraction method which uses EM and GMM models [1] to estimate the background from RGB images. We randomly select 10 sequences from the testing set, and find that our algorithm remarkably outperforms others on all the data. Due to the length limit, we only illustrate the results from one testing sequence in Fig. 4.

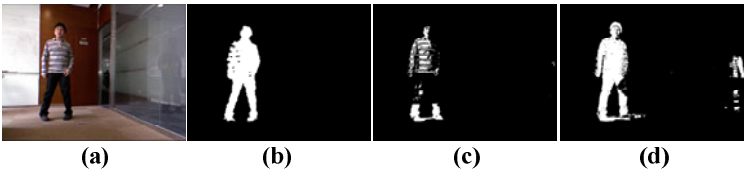

Fig. 4. Performance comparison of human segmentation: (a) original image; (b) result of our algorithm; (c) result of EM based algorithm; (d) result of GMM based algorithm

3.2 Details of Attributes Extraction

Biometrical Attributes. Intuitively, height and shoulder breadth are typical attributes in discriminating different people in surveillance system, which are also noted in some previous works [13, 14, 15]. However, existing height estimation work mainly relies on conventional RGB cameras, and the estimation is not robust to illumination changes. Taking the advantages of RGB-D information in 3D scene establishment, we propose a robust implementation to estimate height and shoulder breadth, and the details are shown as below.

Measuring points are important for height and shoulder breadth measurement. As shown in Fig. 5, we define four measuring points including V_{hh} and V_{hl} as the highest and lowest points for height measuring, V_{sl} and V_{sr} as the leftmost and rightmost points for shoulder breadth measuring. The measurement consists of two steps: detecting the measuring points in individual region and obtaining their location information in real world coordinate.

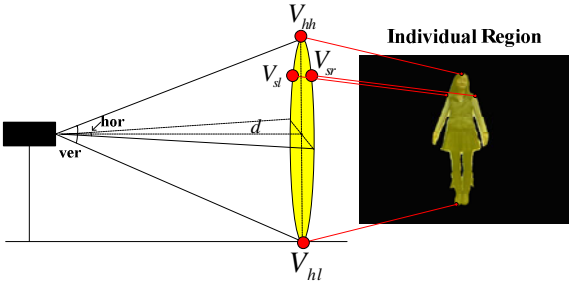

Fig. 5. Automatically measuring height and shoulder breadth

The four points are initialized at the intersection points of bounding rectangle and individual region. The situation of rectification is twofold: having head location or not. The head location is detected by a cascade of Adaboost classifier based on Haar-like features. As only the individual's silhouettes are fed into the classifier, the precision is guaranteed to be high. If head location is detected, V_{hh} will be the highest point of head region. Moving down from the head region, we can get the left shoulder V_{sl} and right shoulder V_{sr}. If there is no head detected, we calculate the depth distribution around V_{hh} to validate whether this point is on the head or not. V_{sl} and V_{sr} are located at the endpoint of the ellipse's horizontal axis, which can be got by the ellipse fitting of trunk. The four points' coordinate figure in pixel coordinates can be reconstructed in real world coordinates through Eq.(2),

$$\begin{cases} x_w = (x_p / X_{res} - \frac{1}{2}) \times d \times \tan(\frac{hor}{2}) \times 2 \\ y_w = (y_p / Y_{res} - \frac{1}{2}) \times d \times \tan(\frac{ver}{2}) \times 2 \end{cases} . \tag{2}$$

Here (x_p, y_p) and (x_w, y_w) are the location of V in pixel coordinate and real world coordinate, X_{res} and Y_{res} are the image width and height resolution, hor and ver are the horizontal and vertical field of view. We suppose the horizontal bisecting line of the camera image is horizontal in the real world as well. The height and shoulder breadth are calculated by $\overline{V}_{hl} - \overline{V}_{hh}$ and $\overline{V}_{sr} - \overline{V}_{sl}$ in world coordinates. $\overline{V}$ is the smoothing value of 5*5 neighborhood of V in individual region.

Appearance Attributes. The appearance attributes of an individual represent the identity of the person [16]. Clothing color and style, skin color and luggage are the most significant attributes for people tracking and identification, as they usually occupy large area in monitor screen and thus are more reliable [5]. The main challenges of extracting appearance attributes are derived from the changes in illumination, pose, and clothing appearance variation [16]. Considering the clothing style is no-rigid deformation with gestures, we just detect it when the people are facing the camera front and standing uprightly.

To compensate the influence caused by the illumination changes, the color of clothing region is transferred from RGB to HSV first. Then the HSV color space is quantified to 24 bins. The quantification method is that the H value is quantified into 6 bins, the S and V values are quantified into 2 bins each. The HSV color histogram is computed to represent the color information.

When two people are dressed up differently but with roughly the same amount of body surface covered with the same colors, they are likely to have similar histogram-based signatures, regardless of how the colors are distributed in space. This is a major limitation of all the holistic models based on histograms because it significantly reduces their discriminability. This issue is addressed here by a self-adapting blocks model covering the whole human body. As shown in Fig. 6, the individual region is divided into 23 blocks covering skin, coat, trousers and luggage. The self-adapting blocks model is structured as below:

1. The external rectangle, individual region and head region have been detected in section 3.1. The head region is considered as one block and the skin color can be obtained by performing a Bayes color model [11] on this block.
2. The body trunk and leg region are detected under the head region and its width equals to shoulder breadth measured before. According to the theory of anthropometry [17], we can divide body trunk region into 14 equal blocks, the top 6 blocks belong to trunk and the button 8 blocks belong to leg.
3. The left and right regions abut against trunk are considered as arms and we divide them into 6 blocks. The two blocks under arms are considered as luggage regions.
4. For each coat, trousers and luggage blocks, we compute the HSV color histogram $H_j(k)$, where $j = 0,\ldots,21$ is block id and $k = 0,\ldots,23$ is the index of histogram bins. The clothing color and style is described by the combination of these blocks. Some examples are shown in Fig. 6.

All blocks' color histograms are indexed into the database. When users want to find people with these attributes, they can choose the color of each block by our interface. If no color is selected for a block, its weight is set to zero. Let $H'_j(k), j = 0,\ldots,22$,

$k = 0,\ldots,23$ is the query condition entered by users. For each people in the database, system will calculate the matching score by Eq.(3),

$$score = (\sum_{j=0}^{22} w_j \sum_{k=0}^{23} \min(H_j(k), H'_j(k))), \quad . \tag{3}$$

Here w_j is the weight of the block j which can be designated by users and $\sum_{j=0}^{22} w_j = 1$. We set the weights of body blocks twice the weights of other blocks by default as the body blocks' color histograms are more significant than others.

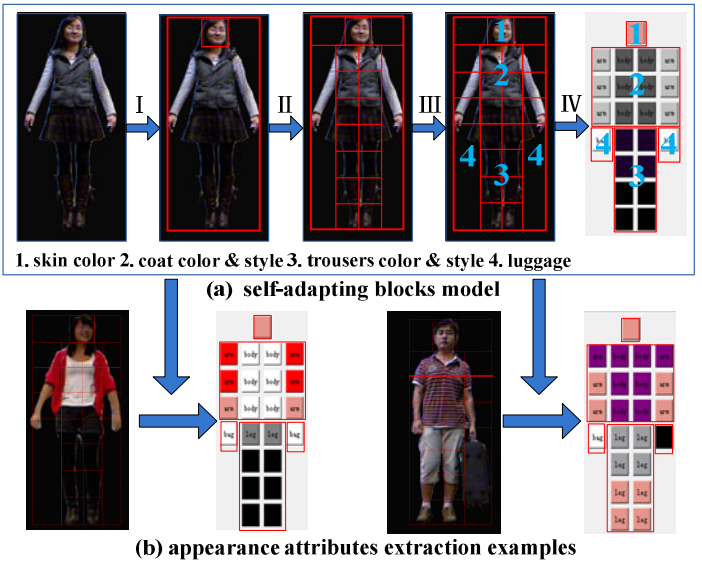

Fig. 6. Appearance attributes extraction. (a) shows the structured of self-adapting blocks model. (b) shows the appearance attributes extracted from two examples which use self-adapting blocks model. (Best seen in color)

Motion Attributes. Motion attributes are related to abnormal behaviors which are mostly concerned by users in visual surveillance [1]. Based on the multiple attributes extracted before, we detect three simple abnormal behaviors, i.e., wandering, running and squatting. The detection is performed as following. Firstly, the biometrical and appearance attributes extracted before are used to track human P_i. Then the stay time T_i of each P_i can be counted, which is used for wandering detection. Meanwhile, the gravity center $G_i(x, y, d, c)$ of each P_i is used to calculate the moving distance D_i in the real world coordinates. So the velocity of P_i is calculated by $S_i = D_i / T_i$, which is used for running detection. At last, the remarkable height changes are believed to be squatting, such as when $ht < \frac{1}{2}\overline{ht}$ suddenly, $\overline{ht}$ is the average height of P_i.

4 Experimental Results

In this section, we comprehensively evaluate the performance of attributes extraction and multi-attribute searching of people. The testing set is captured by a RGB-D camera (also known as Kinect-style camera): it consists of 100 RGB-D video surveillance sequences, captured at three different scenes including meeting room, corridor and entrance, with 50 different persons appearing in it, as shown in Fig. 1. All the people have height distribution of 150~185cm and shoulder breadth distribution of 35~55cm; wearing clothes with different styles and different colors; conduct actions including standing, squatting, jumping, walking, running, rotating and waving hands.

4.1 Performance of Attributes Extraction

Biometrical Attributes. Our system automatically measures the height and shoulder breadth of 38 people from the testing set, whose actual values have been measured manually. The results are shown in Fig. 7. The bias of height is -1.21 cm and standard deviation is 3.18 cm. The negative bias is due to when people standing in normal state, his height is generally lower than standing upright. The bias of shoulder breadth is -0.11 cm and standard deviation is 5.28 cm. The relatively large deviation is due to the interference brought by various poses of people.

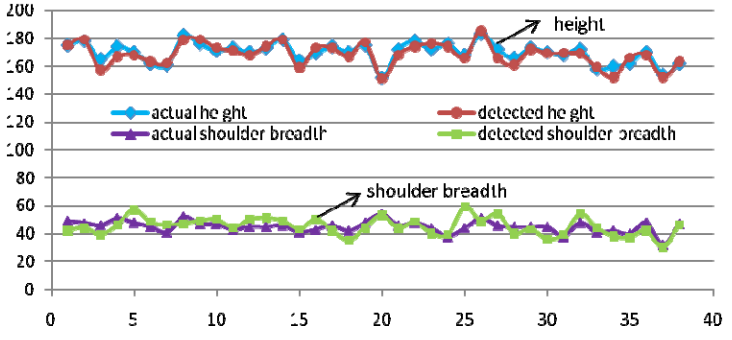

Fig. 7. Performance of height and shoulder breadth measurement. The X-axis is the people id and Y-axis is length unit in cm.

Motion Attributes. The ground-truth of three abnormal behaviors in testing set is labeled manually, and the results are shown in Table.1. Benefitting from the advantages of RGB-D information, wandering detection achieves satisfying performance, running and squatting detection can also meet searching requirements. The reason of the recognition rate of squatting and running is relatively lower than wandering is that the diversity of conduct actions may influence the obtaining of velocity and height changes.

Table 1. Performance of abnormal behavior detection

Abnormal behavior	Recall	Precision
Wandering	0.95	1
Squatting	0.94	0.738
Running	0.86	0.8

4.2 Evaluation of Multi-attribute People Searching

In order to evaluate the effectiveness of multi-attribute based people searching in our IVS system, we conduct a known-person search task which models the situation in which a user has seen a person in the testing set before, but doesn't know where to find it now. We invited 5 users, each user is required to search 10 different known-persons (seen before the searching by users for only once), therefore there are 50 known-person search tasks in all, and each task can be conducted in three rounds. From Fig. 8, we can find: (1) 28 tasks are completed successfully in the first round by appearance attributes including clothing color and style, skin color and luggage; by adding biometrical attributes, the number rises to 37; when all the attributes are used, the number rises to 40; (2) the number of hits rises if the user searches more than one time.

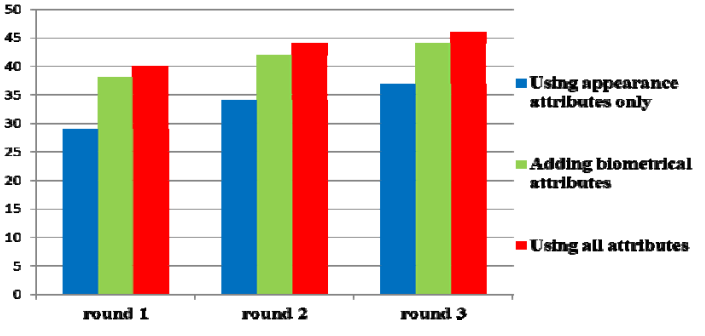

Fig. 8. Performance of 50 known-person search tasks conducted on our system by multi-attribute searching. The X-axis is search times and Y-axis is the number of hits.

5 Conclusions and Future Work

We innovatively bring RGB-D information into IVS system to assist multi-attribute based people searching. Taking the advantages of RGB-D information in human segmentation and 3D scene establishment, we have extracted biometrical attributes including height and shoulder breadth; appearance attributes including clothing color and style, skin color and luggage information; and also motion attributes including the detection of squatting, running and wandering. The comprehensive evaluations on our IVS system demonstrate the effectiveness of the extracted multiple attributes and their successful application in multi-attribute based people searching.

Our attempts in this paper indicate that RGB-D information has promising potential in IVS field. In the future, we will investigate how to further combine RGB and Depth

information in a multiple graph framework [18][19] to discover more useful and challenging attributes in surveillance environments.

Acknowledgments. This work is supported by the National High Technology and Research Development Program of China (863 Program, 2009AA01A403); National Nature Science Foundation of China (61100087); Beijing Natural Science Foundation (4112055); Beijing New Star Project on Science & Technology (2007B071); Co-building Program of Beijing Municipal Education Commission.

References

1. Hu, W.M., Tan, T.N., Wang, L., Maybank, S.: A Survey on Visual Surveillance of Object Motion and Behaviors. IEEE Trans. on SMC (2004)
2. RGBD Projects, http://ils.intel-research.net/projects/rgb
3. Henry, P., Krainin, M., Herbst, E., Ren, X.F., Fox, D.R.-D.: Mapping: Using Depth Cameras for Dense 3D Modeling of Indoor Environments. In: ISER (2010)
4. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-Time Human Pose Recognition in Parts from a Single Depth Image. In: CVPR (2011)
5. Vaquero, D., Feris, R., Tran, D., Brown, L., Hampapur, A., Turk, M.: Attribute-Based People Search in Surveillance Environments. In: WACV (2009)
6. Siddiquie, B., Feris, R.S., Davis, L.S.: Image Ranking and Retrieval based on Multi-Attribute Queries. In: CVPR (2011)
7. Lin, C.H., Chen, A.L.P.: Indexing and Matching Multiple-Attribute Strings for Efficient Multimedia Query Processing. IEEE Trans. on Multimedia 8(2), 408–411 (2006)
8. PrimeSense, http://www.primesense.com/
9. OpenNI, http://www.openni.org/
10. Xia, L., Chen, C.C., Aggarwal, J.K.: Human Detection Using Depth Information by Kinect. In: International Workshop on Human Activity Understanding from 3D Data in conjunction with CVPR, HAU3D (2011)
11. Jones, M.J., Rehg, J.M.: Statistical Color Models with Application to Skin Detection. IJCV 46(1), 81–96 (2002)
12. Lai, K., Bo, L.F., Ren, X., Fox, D.: A Large-Scale Hierarchical Multi-View RGB-D Object Dataset. In: ICRA (2011)
13. Kispal, I., Jeges, E.: Human Height Estimation Using a Calibrated Camera. In: CVPR (2008)
14. Gallagher, A., Chen, T.: Jointly Estimating Demographics and Height with a Calibrated Camera. In: ICCV (2009)
15. Madden, C., Piccardi, M.: Height Measurement as a Session-based Biometric for People Matching Across Disjoint Camera Views. In: IAPR (2005)
16. Doretto, G., Sebastian, T., Tu, P., Rittscher, J.: Appearance-based Person Reidentification in Camera Networks: Problem Overview and Current Approaches. Journal of Ambient Intelligence and Humanized Computing 2(2), 127–151 (2010)
17. Anthropometry, http://personal.cityu.edu.hk/~meachan/Online%20Anthropometry
18. Wang, M., Hua, X.S., Hong, R.C., Tang, J.H., Qi, G.J., Song, Y.: Unified Video Annotation via Multi-Graph Learning. IEEE Trans. on Circuits and Systems for Video Technology 19(5) (2009)
19. Xia, T., Tao, D.C., Mei, T., Zhang, Y.D.: MultiView Spectral Embedding. IEEE Trans. on Systems, Man, and Cybernetics: Part B 40(6), 1438–1446 (2010)