ステレオカメラによる人の位置検出および動作分析技術

Stereo Vision-Based Person Location and Behavior Analysis

范　聖印*　　王　鑫*　　王　千*　　乔　刚*　　诸　加丹*

Shengyin FAN　　Xin WANG　　Qian WANG　　Gang QIAO　　Jiadan ZHU

要　旨

人の位置，移動経路および動作は非常に重要な情報であり，これらは人の位置に関連するサービスを提供するための基礎的材料となる．しかし，従来型の2D情報を取得するだけのカメラでは，実際の位置，移動経路，正確な動作を提供することができない．リコーは，通常のRGB情報以外に奥行き情報を提供することができるカメラを開発している．このカメラは，例えばMicrosoftのKinectのようなアクティブライト方式のステレオカメラに比較して，はるかに大きい視野を有している．

我々は，ステレオカメラを用い，人の位置を検出し，動作（立つ，歩く，座るなど）を分析することで，省エネルギー，監視カメラおよび室内の人流解析などのような人の位置関連サービスに向けた情報提供システムを目指す．3D情報，高さ検出による人の姿勢認識およびRandom Forests法による人の向き情報検出などによるアルゴリズムを適用することで人の特定，追跡，位置検出などを可能とするステレオビジョンを提案する．

我々が開発したソリューションでは，1つまたは複数のステレオカメラを用いてリアルタイムに人の位置情報と動作情報を提供することが可能である．

ABSTRACT

Information on a persons' location, trajectory and behavior is very important as it forms the foundation of human-centered services. Traditional cameras capture 2D information from 3D spaces, but fail to provide real locations, movement trajectories and accurate behavior. Ricoh is developing a binocular camera that provides the disparity/depth information output in addition to the normal RGB information. Its working range is much bigger than active light-based depth sensors such as those of the Microsoft Kinect camera.

Our goal is to use a binocular camera to locate people, analyze their behavior (standing, walking, sitting) and recognize their direction to provide human-centered services such as energy efficiency in the office, intelligent surveillance, location-based marketing. We propose using stereo vision-based algorithms, which include an algorithm to detect, track and locate a person, as well as recognizing posture by height distribution, and direction by Random Forests.

The developed solution supports real time location and behavior analysis using one binocular camera and multiple binocular cameras.

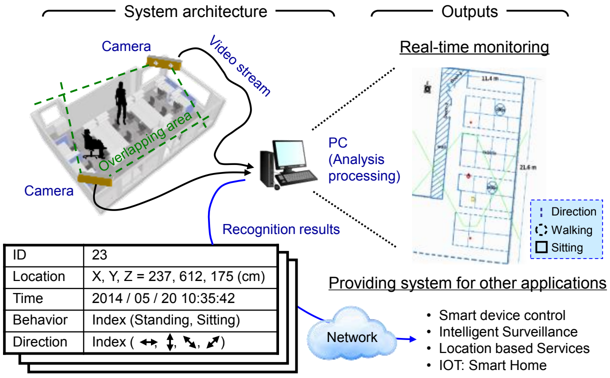
*　リコーソフトウェア研究所（北京）有限公司

Ricoh Software Research Center (Beijing) Co., Ltd.

1. Introduction

A person’s location, trajectory, and behavior are very important, as they are the basis of human-centered services. This is a challenging problem in computer vision. A traditional camera captures 2D information from a 3D space, and cannot provide a person’s real location and trajectory, making it difficult to accurately recognize behavior.

Ricoh is developing a binocular camera whose video stream includes an RGB/grayscale stream and a depth/disparity stream, which can display actual 3D scenarios much better. Its working range is much bigger than that of an active light-based camera, such as the Microsoft Kinect, meaning it could be deployed in larger areas.



ID	23
Location	X, Y, Z = 237, 612, 175 (cm)
Time	2014 / 05 / 20 10:35:42
Behavior	Index (Standing, Sitting)
Direction	Index (↔, ↕, ⤡, ⤢)

Fig. 1 Schematic Diagram.

The schematic diagram of application scenarios is shown in Fig. 1. It supports multiple binocular cameras (two cameras in the diagram). By simultaneously processing the video streams of different binocular cameras, the person’s tracking ID, location, behavior, direction and timestamp are calculated, thus achieving real time monitoring of a person's location and behavior. Finally, it supports mobile applications, which enable the remote control of energy efficiency in the office, intelligent surveillance, location-based marketing, and Internet of Things applications (smart homes, the elderly taking care of themselves at home, etc.).

The paper is organized as follows: Section 2 introduces the current state of this technology, including the traditional, camera-based solution and stereo vision-based solution. Section 3 presents the technology solution in detail, specifically the following parts: Ricoh’s high dimensional binocular camera and its output, 3D foreground point cloud projection-based person detection and tracking, temporal height distribution change accumulation-based posture recognition, Random Forest-based direction recognition, and observable feature-based fusion. Next, Section 4 discusses the current results and known issues. Finally, Section 5 concludes the paper with a brief summary and presents an overview of future research.

2. Related Work

Verient[1] proposes an intelligent video analysis solution for retail, supermarket and security applications. Intel[2] proposes a similar solution for shopping malls in Beijing. Both of them use RGB video streams for detection and tracking, behavior analysis, crowd flow and density of people. In both scenarios the results for crowded areas were unsatisfactory.

Microsoft[3] and PrimeSense[4] use an active light device to generate disparity/depth information. Applications mainly focus on the interaction between humans and computers in gaming. Their person detection, tracking and behavior analysis showed satisfactory results, but the working range was limited to several meters, making it difficult to use in office and supermarket settings, for security, etc.

Point Grey[5] and TYZX[6] both sell binocular cameras, but Point Grey only provides the software development kit (SDK) for capturing RGB and depth images. TYZX provides solutions for detecting, tracking, gesture recognition, car detection, etc., but the image is low resolution and the results proved unsatisfactory in crowded scenarios.

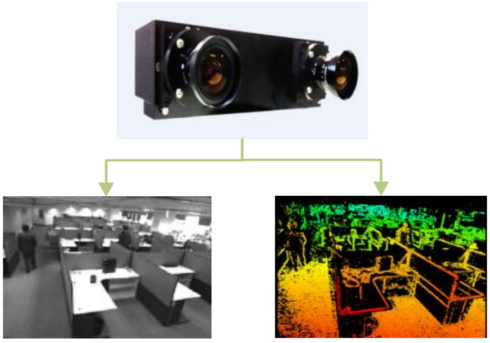
3. Technology Solution

3-1 Overview

In this section we describe the detailed technology solution proposed for supporting the application scenarios shown in Figure 1. It comprises the following five parts: the Ricoh binocular camera and its output, 3D foreground point cloud projection-based person detection and tracking by a single binocular camera, temporal height distribution variation accumulation-based posture recognition using a single binocular camera, Random Forests-based direction recognition by single binocular camera, and observable feature-based fusion for supporting multiple binocular cameras.

3-2 High Dimensional Binocular Camera and Its Output

The binocular camera is a new 3D sensor for large range stereo vision sensing. Few of these products are available on the market apart from the TYZX G3 EVS and the Point Grey Bumblebee 2. Compared with them, Ricoh's binocular camera's resolution is much higher (1280x960).



Grayscale Image

Depth Image

Fig. 2 Binocular camera and its output.

Fig. 2 illustrates the binocular camera and its output. It outputs a grayscale image and a depth image simultaneously. The camera's interface is GigE Vision, which is easy to deploy and use. In the depth image, different colors indicate different distances.

3-3 3D Foreground Point Cloud Projection-based Detection and Tracking

In this section, we propose a single binocular camera-based detection and tracking method. An overview is shown in Fig. 3. The input is the binocular camera's grayscale stream and depth stream. The first processing is foreground extraction by background modeling. The second processing is using depth to undergo two transformations for foreground to build up 3D points cloud. The third processing is constructing three different plan view map. The final processing is Plan view-based person detection and tracking.

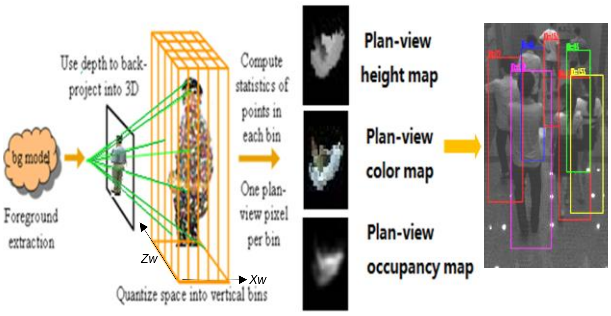

bg model
Foreground extraction
Use depth to back-project into 3D
Zw
Xw
Quantize space into vertical bins
Compute statistics of points in each bin
One plan-view pixel per bin
Plan-view height map
Plan-view color map
Plan-view occupancy map

Fig. 3 Projection-based detection and tracking.

The grayscale image and depth streams are used to conduct background modeling (dynamic modeling using the Gaussian mixture model). To construct the plan view map, the foreground pixels must undergo two transformations from 2D to 3D. The first transformation is from the image coordinate (2D coordinate) to the camera coordinate (3D coordinate), which is accomplished by using Formula 1.

$$X = (x_l - c_x) * Z/f_x$$
$$Y = (y_l - c_y) * Z/f_y \qquad (1)$$

In it, x_l, y_l are the image coordinate values (u, v). c_x, c_y, f_x and f_y are the intrinsic parameters of the binocular camera, Z is the depth value obtained from the depth/disparity of the image, and X, Y are the pixel coordinate values in the camera coordinates.

The second transformation is from the camera coordinate (3D coordinate) to the world coordinate (3D coordinate), which is obtained by using Formula 2.

$$\begin{bmatrix} X_W \\ Y_W \\ Z_W \end{bmatrix} = R \begin{bmatrix} X_C \\ Y_C \\ Z_C \end{bmatrix} + T \tag{2}$$

In it, R is the rotation matrix and T is the translation vector. They are calculated by the camera calibration process. (X_C, Y_C, Z_C) are the camera coordinate values, while (X_W, Y_W, Z_W) are the transformed results which are the uniform world coordinate value of the foreground pixel.

The 3D point cloud was obtained from the projection, after which we had to choose a resolution at which to quantize the world space into vertical bins. We used vertical bins that intersect the X_W Z_W ground-level-plane so as to divide it into a square grid with a resolution of 4 cm/pixel. Each bin's pixel number is calculated to generate the occupancy map, each bin's max height is calculated to generate the height map, and each bin's main color is calculated to generate the color map. This completes the plan view map construction.

Fig. 4 shows the overview of this method.

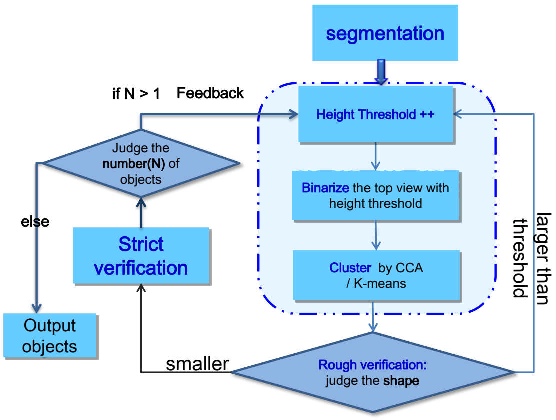


Fig. 4 Segmentation and hybrid verification.

The key contribution of this method is a newly proposed segmentation and hybrid verification based-person detection.

As for the segmentation, we performed an iterative segmentation on the plan view height map. The main steps of each iteration are: increase the height threshold, use the height threshold for the plan view height map binary, and segment the binary image using a general spatial clustering method (such as connected components analysis or K-mean clustering).

The important issue is determining the termination criterion to differentiate between individuals so that the segmentation iteration can stop. The most recently segmented parts for each iteration go to the hybrid verification stage. The hybrid verification stage consists of two types of criterion: rough criterion and strict criterion. They use motion vectors and the object's height distribution feature on the plan views.

Rough criterion is defined as the shape of the segment parts that could be compared with a standard shape threshold to judge whether there is more than one person. If its shape is larger than the threshold, the segmentation iteration continues. If neither of these shapes can be used, a strict verification is used. When objects are very close and obstructed, the shape on the plan view is smaller than it is for an individual person, meaning a progressive verification is needed after the rough verification.

Strict criterion is defined as the scenario where motion vectors within the segment block are clustered. If there is more than one cluster, they are separated and the segmentation iteration is terminated. If necessary, a confidence score is computed by matching the height distribution with the single person height distribution feature. If the confidence score is larger than a given score, the segment part is output and the segmentation iteration is terminated, otherwise, the iteration continues. Fig. 5 shows the iterative segmentation and verification results of three trials. As more time iterations are performed, more persons are detected.

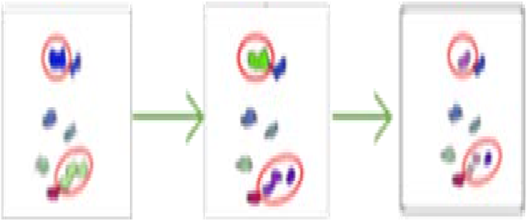


Fig. 5 Results of iterative segmentation and verification.

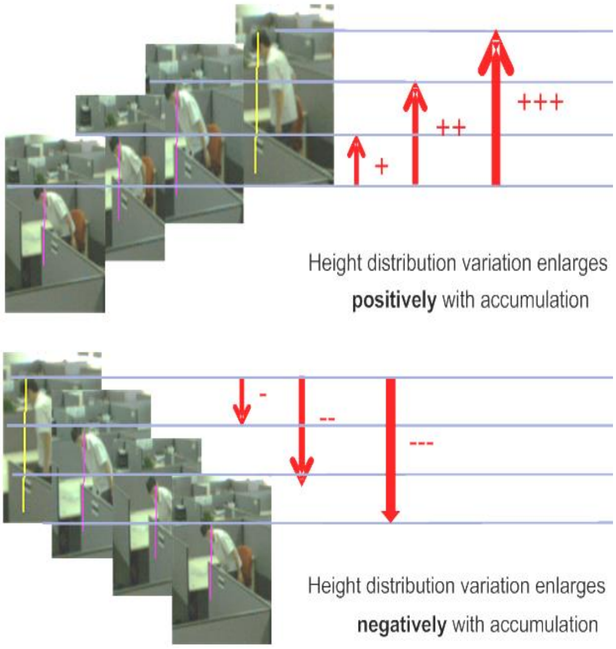
As for tracking, Kalman filtering[7] is used to track patterns of plan-view height, color and occupancy statistics over time.

3-4 Temporal Height Distribution Variation Accumulation-based Posture Recognition

In this section, the single binocular camera-based posture recognition method is proposed, which is based on detection and tracking results. After we track a person, we get the height distribution/map of that person frame by frame. The key contribution uses the sequence of temporal height distribution/map as input and accumulates/sums its variations/differences. Different postures are judged on the changing trend and degree of variation.

Fig. 6 Sequence of temporal height distribution/map.

Fig. 6 illustrates the sequence of temporal height distribution/map, which comprises several frames of the plan view height map.



+
++
+++
Height distribution variation enlarges positively with accumulation
-
--

Height distribution variation enlarges negatively with accumulation

Fig. 7 Trend of variation for different actions.

Fig. 7 illustrates two different accumulation situations. In the first situation, the person is standing up and height distribution variation increases with accumulation. In the second situation the person is sitting down and height distribution variation decreases with accumulation. By using temporal height distribution accumulation as the basis for judgment, it enhances the stability of the results and reduces false positives.

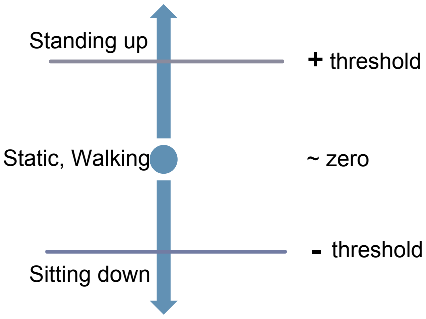


Fig. 8 Action is judged in accordance with threshold.

The posture judgment is based on whether an action has taken place. Fig. 8 displays how we used different thresholds for action detection. If the change trend of accumulation is positive and is bigger than the positive threshold, it means that a standing up action has occurred and was detected and the person is standing. If the change trend of accumulation is negative and is less than the negative threshold, it means that a sitting down action was detected and the person is sitting. If the change trend is close to zero, it means no action has occurred and indicates that the person maintained their previous posture or is walking.

3-5 Random Forests-based Direction Recognition

In this section, we propose a single binocular camera-based direction recognition method in which the key contribution is the person's head/shoulder height distribution map.

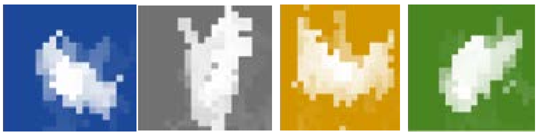


Fig. 9 Head/shoulder height distribution map.

Fig. 9 illustrates the head/shoulder height distribution maps of four different people in four different direction situations. The head/shoulder height distribution map is extracted from the person's plan view height distribution map after locating the head position of the person.

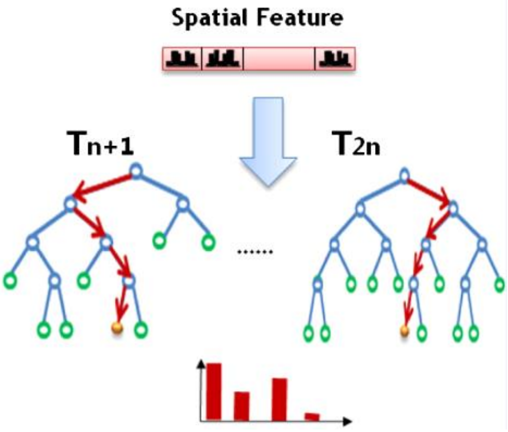


Fig. 10 Random Forests-based direction recognition.

Fig. 10 shows how we used a distribution map as a spatial feature, and then used Random Forests[8] to train the classifier. After training, we used that system to judge the person's direction based on the person's head/shoulder distribution map.

3-6 Weighted Observable Features-based Fusion

In this section, we propose a multiple binocular camera-based fusion method, which enhances person detection and tracking by using different binocular cameras. The key contribution is a new feature called weighted observable features.

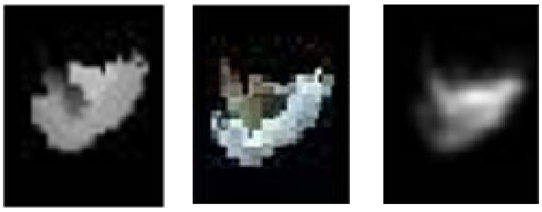


Fig. 11 Features: height, color, occupancy map.

Observable features mean the plan view height map, color map, and occupancy map of a detected and tracked person obtained from different binocular cameras. Fig. 11 shows some examples of them. From left to right, they are height, color and occupancy.

When fusing the results of the same person from different binocular cameras, the first step is to judge which binocular camera detected and tracked the person more accurately so that this camera's higher score can be weighed more heavily. We also propose the addition of the following two factors:

The first factor is the distance between the binocular camera and the person. Errors in measuring depth are an intrinsic problem of binocular cameras, which perform well at short distances, but produce large depth measurement errors at larger distances. When two cameras are used to track the same person, the binocular camera closer to the person produces a more reliable result.

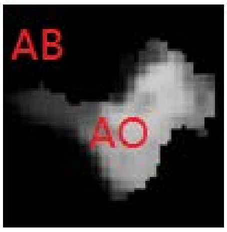


Fig. 12 Visibility degree definition.

The second factor is the visibility degree of the person to the binocular camera. Fig. 12 shows an example of this. The visibility degree is used to solve the problem of how clearly a person is exposed to the binocular camera, which is defined on an occupancy map. The concept of visibility degree is defined as:

$$VisDegree = \frac{AO}{AB} \tag{3}$$

AO is the area of the occupancy map and within the bounding box where the intensity is not black, and *AB* is the area of the bounding box. The higher the ratio *VisDegree* is, the more visible the person is to that camera.

By considering the distance and visibility degrees, it is easy to decide what weight should be given to different binocular cameras and how the weight for observable features is determined. Using the weighted observable features of each person and that person's position, the judgment of whether to fuse the results of persons from different binocular cameras into the results of one person has to be performed.

4. Results

4-1 Detection and Tracking

We evaluated the accuracy of person detection and tracking by using a single binocular camera in different scenarios in the office. For all of the scenarios, the lighting used conforms to normal standards: day light with no major lighting changes. The person's walking speed is normal and he/she does not run. The office map is shown in Fig. 13. Two binocular cameras were deployed in the office, which are represented with red points.

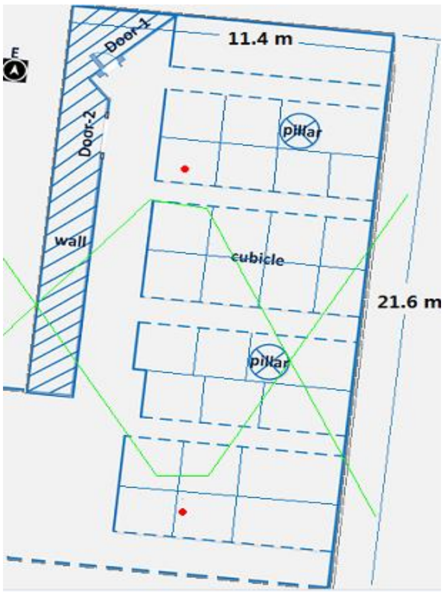


Fig. 13 Office map for evaluation.

The different scenarios take into consideration the number of people and the density of the crowd. We employed four scenarios: 2 persons, 4 persons, 6 persons and a crowded office. In the first three scenarios, each person could walk around and choose his/her seat at random. In the final scenario, people were kept within certain parameters.

Table 1 summarizes the results obtained by using only one binocular camera. In the crowded scenario, the results were less satisfactory due to the obstruction of people. Fig. 14 shows the detection and tracking results for the crowded scenario.

Table 1 Person detection and tracking evaluation.

Testing Scenarios	Detection Rate	False Alarm Rate
2 persons	89%	3.3%
4 persons	87%	3.0%
6 persons	84%	3.7%
Crowded	71%	5.6%

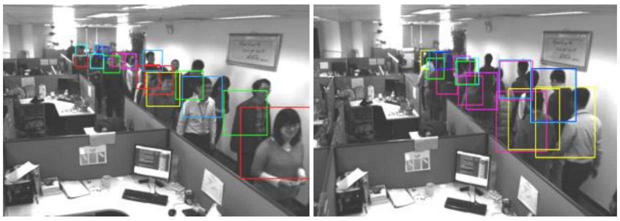


Fig. 14 Crowded scenario detection and tracking results.

Furthermore, we also evaluated the location accuracy for individuals and found that it is mainly dependent on the person's distance to the binocular camera. On average, the accuracy of location is within 20 cm when the camera is between 2 meters to 10 meters away. When the camera is between 10 meters to 20 meters away the location accuracy drops to 50 cm. When the camera is more than 20 meters away the accuracy of the location degrease more rapidly.

4-2 Recognition of Three Postures

As for the accuracy in posture recognition, we mainly evaluated the first three scenarios described in section 4-1, because posture is not needed in real-life applications within a crowded environment. Fig. 15 shows the posture recognition results for a two-person scenario where one person is walking and the other is sitting.

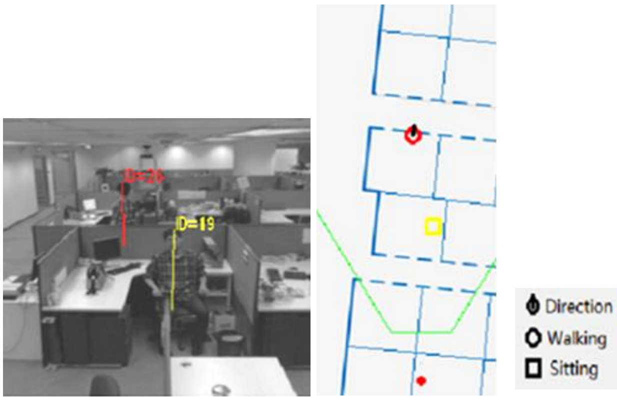


Fig. 15 Posture recognition results.

For sitting, walking and standing recognition, our method achieves a 90% recognition rate with a 5.6% false alarm rate.

4-3 Four Direction Recognition

As for the accuracy of direction recognition, we also focused on evaluating the first three scenarios described in section 4-1. Considering direction could be very useful in a crowded scenario, so we prepared a scenario with five people within a limited area. Fig. 16 shows the direction recognition results for this scenario.

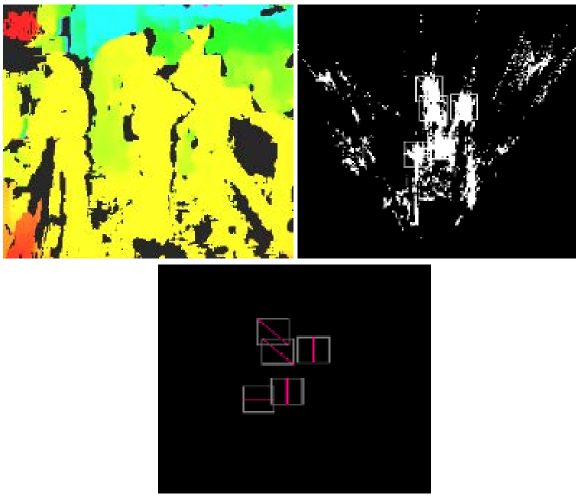


Fig. 16 Direction recognition results.

In Fig. 16, the top left image is the depth, the top right image is the height map and the bottom image is the results. The directions of all five people were recognized correctly.

For the three, non-crowded scenarios described in section 4-1 the recognition for all four directions were 75%, with a false alarm rate of 7.5%. In the crowded scenario we achieved 67% recognition for all four directions, with a false alarm rate of 8.5%.

4-4 Fusion Results

We evaluated the accuracy of person detection and tracking using two binocular cameras in accordance with the four scenarios described in section 4-1.

Table 2 summarizes the results obtained by using two binocular cameras. Overall, using the fusion results obtained by two separate cameras yields better results than using a single binocular camera. Furthermore, the results obtained from crowded scenarios are always less satisfactory than those obtained in the other scenarios.

Table 2 Evaluation results of using two cameras.

Testing Scenarios	Detection Rate	False Alarm Rate
2 persons	97%	4.1%
4 persons	93%	3.7%
6 persons	91%	3.7%
Crowded	77%	6.1%

Localization accuracy depends mainly on the distance to the binocular camera, indicating that for the same range area, multiple binocular cameras achieve better localization accuracy than a single binocular camera.

4-5 Known Issues Summary

After conducting the above evaluation we discovered some known issues. The first is how to enhance the detection and tracking performance for crowded scenarios. Our initial thoughts on this were to adaptively adjust the detection and tracking match threshold. The second know issue is how to expand the recognized postures–which are both simple and rare–and how to propose a new algorithm for them. The third is to expand the directional recognition to include eight directions, as four directions are insufficient. One way to achieve this is to consider the color and height map together. The final known issue is the fusion of posture, direction, etc., and what features could be implemented to address obstructions.

5. Conclusion

In this paper we proposed algorithms and developed a solution that allows us to use one and multiple binocular cameras to detect, track and locate individuals, analyze their behavior (standing, walking, and sitting) and recognize their direction. The proposed solution achieves results in actual tests and shows potential for use in office energy efficiency, intelligent surveillance, location-based marketing, etc. This is the first attempt at using binocular cameras and this field warrants further research. In the future, we will focus on surveillance, user experience and expectations, and further investigate the potential advantages of the binocular camera, enhance algorithms by combining depths and RGB information, and find suitable marketing solutions using the current technology available.

References

1) Verient: Video & Situation Intelligence, http://www.verint.com/solutions/video-situation-intelligence/index.html (accessed 2014-07-06).

2) Intel: Intel Retail Innovations, http://www.intel.com/content/www/us/en/retail/retail-innovations.html (accessed 2014-07-07).

3) Microsoft: KINECT for Windows, http://www.microsoft.com/en-us/kinectforwindows/ (accessed 2014-07-07).

4) PrimeSense: 3D Sensing Company, http://en.wikipedia.org/wiki/PrimeSense (accessed 2014-07-07).

5) TYZX: Person and Gesture Tracking with Smart Stereo Cameras, http://www.tyzx.com/PDFs/TyzxSPIE2008Dist.pdf (accessed 2014-07-07).

6) Point Grey: Bumblebee2, http://ww2.ptgrey.com/stereo-vision/bumblebee-2 (accessed 2014-07-07).

7) C. K. Chui, G. Chen: Kalman Filtering with Real Time Applications, Springer Science & Business Media (December 1, 2008).

8) L. Breiman: Random Forests, *Machine Learning*, Vol. 45, Issue 1, pp. 5-32 (October 2001).