ISZOT: Extracting Zone-based Trajectory of Shopper for Marketing Analysis

Yuan Li∗, Masanori Miyoshi∗, Shun'ichi Kaneko†, Hironobu Fujiyoshi‡,
Toshiki Etchuya†, Hiroyuki Nara†, Kotomi Watanabe§

∗ Hitachi Research Lab., †Hokkaido University , ‡Chubu University,
§Consumers Co-operative Sapporo

E-mail: yuan.li.yw@hitachi.com

Abstract

In marketing, sales are evaluated and improved by analyzing the purchasing behavior of shoppers, such as when, where and what they purchased. However, most analysis is based on POS (Point of Sale) data or on surveys carried out by human beings. This is not only costly and time consuming, but also lacks visual information on shoppers and their purchasing process. In order to overcome these problems, we propose a new and efficient surveillance system, ISZOT: Image based sensing system for analyzing Shoppers' purchasing behavior based on ZOne Trajectories. This surveillance system uses images from surveillance cameras to extract a shopper's trajectory in order to analyze their purchasing behavior for different marketing purposes. In this paper, ISZOT and its essential technologies will be introduced. In ISZOT, considering the accuracy and performance of image sensing process, a new human detection approach using a Smart Window Transform (hereafter SWT) and an edged-based classifier, and shoppers' trajectories extraction approach is proposed. Experiments to evaluate the effectiveness of a store's layout based on shoppers' trajectories in an observed area are implemented with the cooperation of the Consumer Co-operative Sapporo, one of the largest chain stores in Hokkaido, Japan. The results show that our proposed system is capable and effective in analyzing shoppers' purchasing behaviors for different marketing usages.

1 Overall of ISZOT

ISZOT is short for Image based sensing system for analyzing Shoppers' purchasing behavior based on ZOne Trajectories. It contains four main parts: human detection, trajectory extraction, zone trajectory extraction and behavior analysis. Figure 1 shows a block diagram of this ISZOT system. In the following, we describe each step briefly.

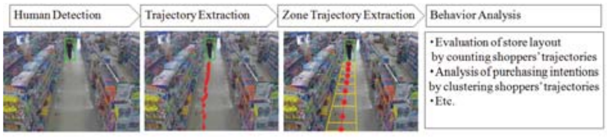

Human Detection
Trajectory Extraction
Zone Trajectory Extraction
Behavior Analysis
•Evaluation of store layout by counting shoppers' trajectories
•Analysis of purchasing intentions by clustering shoppers' trajectories
•Etc.

図 1 Block diagram of the ISZOT system

At the first step, human detection using a SWT and an edge-based classifier is implemented. A SWT is a new approach to detect people in the real world coordinate, which is robust to distortion caused by different camera settings. A novel Joint-HOG method is performed with 2-stage Adaboost to classify human beings. This method can automatically capture human shapes like shoulders, legs, and arms, symmetrically and continually, and has a high classification rate. After classification, all results are integrated by applying a Mean-shift algorithm in the world coordinate to obtain the final human detection results.

At the second step, the human trajectory is extracted based on the human detection results. This is carried out by a simple human tracing algorithm based on Nearest Neighbor.

At the third step, the human zone trajectories are extracted by combining the shoppers' trajectories with the display zones of goods in the shops. This can provide semantic labels to a shopper's behavior.

At the last step, shoppers' behavior is analyzed based on the extracted trajectories. Experiments in a real supermarket are carried out.

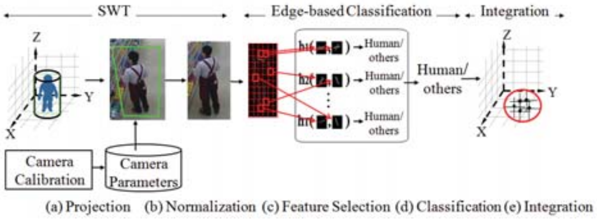


図 2 Overall follow of human detection

2 Human Detection

In ISZOT, human detection is the most important task to obtain a shopper's trajectory. In this paper, we proposed a new training based human detection method based on SWT and edge-based features. Figure 2 shows the overall flow of the proposed approach. First, for any observed position in the world coordinate, the SWT is performed to get a detection window with the best shot of a person and transfer it into the image coordinate. Then a normalization is performed to transfer this detection window into a rectangle to fit a human classifier. Later, an edge-based classifier is used to extract the local features inside the normalized window and classify whether it is human or not. This procedure is repeated until all observed positions in the world coordinate are classified. Lastly, all the classified results are presented and integrated in the world coordinate.

2.1 Smart Window Transform

In human detection, distorted or tilted human forms are difficult to classify. These distortions are due not only to the lens distortion but also camera settings such as large pitch or roll angles and so on. A human silhouette extracted from rectangular detection windows is different from the training samples. Although we can train these distortion samples in advance, it is costly and unrealistic to collect all training samples for each individual camera setting or detection situation. To overcome this problem, a new method called Smart Window Transform (SWT) [1] is proposed and implemented. This method generates a detection window in the world coordinate and transfers it into a rectangle for classification. SWT is robust to camera setting and contains less human distortion. Figure 3 shows the SWT model when detection is implemented at position A. A detection window is generated in the world coordinate. As shown in Figure 3, the person is illustrated as a cylinder. D is the height and W is the width of the person. For a given camera m, a cross-section of the cylinder is generated with bottom center located at A and upper and lower edges parallel to the X axis of camera m. This cross-section is called a Smart Window (hereafter SW) in this paper. As shown in Figure 3, the four vertexes of the SW are presented as $\{{S_j}^m | j = 1, 2, 3, 4\}$. For camera n, a SW can also be generated and presented as $\{{S_j}^n | j = 1, 2, 3, 4\}$. This SW always faces the camera to obtain the best shot of the person for classification.

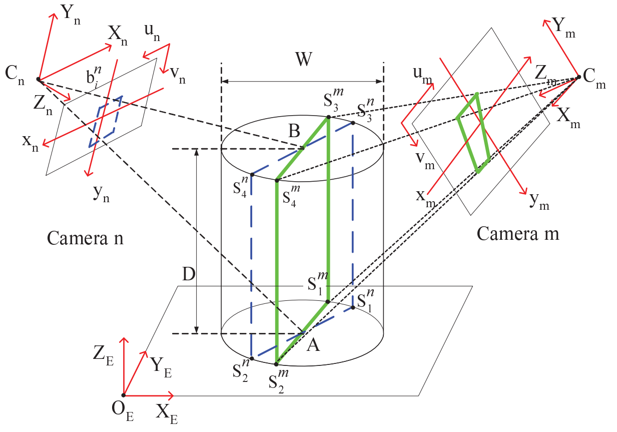


図 3 Smart Window Transform model

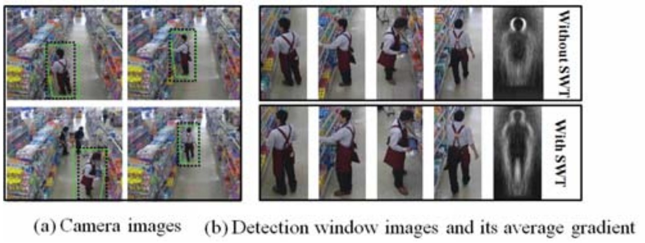
A SWT is then performed to extract the human best shot from the images and normalize it to suit a classifier. Camera m is used for the following explanation. Firstly, the generated SW is transformed from the world coordinate, $O_E - X_E Y_E Z_E$, to the camera coordinate, $C_m - X_m Y_m Z_m$. In Figure 3, the upper and bottom edges of the SW are parallel to the X_m axis of camera m. The intersections of the bottom edge of the SW and the cylinder, ${S_1}^m$, ${S_2}^m$ can be computed from A as:

$$ {S_1}^m, {S_2}^m = M_m A \pm \begin{bmatrix} \frac{1}{2}W & 0 & 0 \end{bmatrix}^T \qquad (1) $$

where, M_m is the transformation matrix, which is the product of a translation and the rotation matrix from the camera extrinsic parameters. Similarly, the intersections of the upper edge of the SW and the cylinder can be computed from B, where $B = A + \begin{bmatrix} 0 & 0 & D \end{bmatrix}^T$. Secondly, the SW is transformed from the camera coordinates to the image coordinates, $o_m - u_m v_m$, as:

$$ k{s_j}^m = H_m P_m {S_j}^m, \ j = 1, 2, 3, 4 \qquad (2) $$

where, k is a scalar, P_m is the perspective projection matrix from the camera coordinate to the image plane $c_m - x_m y_m$, and H_m is the transformation matrix from the image plane to the image coordinate, which is com-



(a) Camera images (b) Detection window images and its average gradient

図 4 Features of Smart Window Transform

posed by camera intrinsic parameters. Camera intrinsic and extrinsic parameters are both achieved by the Tsai camera calibration method [2].

Finally, the SW is normalized to a rectangle adapted to a classifier. Perspective projection is performed for normalization as:

$$u_j = \frac{c_{00}{u_j}^m + c_{01}{v_j}^m + c_{02}}{c_{20}{u_j}^m + c_{21}{v_j}^m + c_{22}}, j = 1,2,3,4$$

$$v_j = \frac{c_{10}{u_j}^m + c_{11}{v_j}^m + c_{12}}{c_{20}{u_j}^m + c_{21}{v_j}^m + c_{22}}, j = 1,2,3,4 \qquad (3)$$

where $({u_j}^m, {v_j}^m)$ is the coordinate of the SW$\{{s_j}^m | j = 1,2,3,4\}$. (u_j, v_j) is the corresponding coordination of the destination rectangular vertex. After these transformation, a new detection window is generated for human classification.

Figure 4 shows the features of the SWT. There are three main features:

(1) The SWT is robust for camera settings. At the first step, we eliminate radial distortion of the image. Then SWT is applied. In Figure 4 (a), the solid detection window is the result of SWT and the dashed window is the result of traditional method. For each camera, the detection windows produced by the SWT always face the camera and contain the best shot of the person's silhouette. The upper and bottom edges are parallel to the X axis of the camera. As shown in Figure 4 (a), due to the camera setting, human forms are distorted or tilted when he is closed to the camera, such as the down left or down right areas in the image. These distortions are decreased when he is away from the camera, such as the middle or upper areas in the image. Figure 4 (b) shows the detection window images with or without applying SWT. In the detection window images without applying SWT, distorted or tilted human forms exist, especially in the areas closed to camera. On the contrary, in the detection window images with applying SWT, these distortions in the human's forms are eliminated. These detection windows with applying SWT are robust to camera setting.

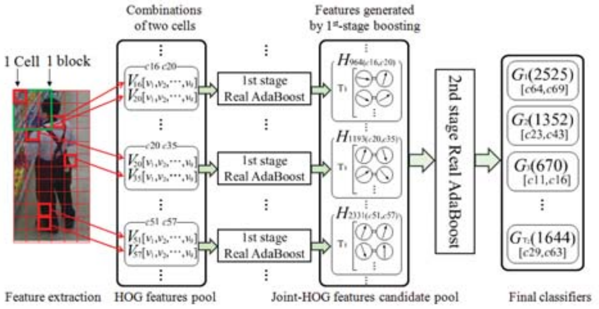

Combinations of two cells
Features generated by 1st-stage boosting
1 Cell
1 block
c16 c20
$V_{16}[v_1, v_2, \cdots, v_9]$
$V_{20}[v_1, v_2, \cdots, v_9]$
c20 c35
$V_{20}[v_1, v_2, \cdots, v_9]$
$V_{35}[v_1, v_2, \cdots, v_9]$
c51 c57
$V_{51}[v_1, v_2, \cdots, v_9]$
$V_{57}[v_1, v_2, \cdots, v_9]$
1st stage Real AdaBoost
1st stage Real AdaBoost
1st stage Real AdaBoost
$H_{964(c16,c20)}$
$H_{1193(c20,c35)}$
$H_{2331(c51,c57)}$
T_1
2nd stage Real AdaBoost
$G_1(2525)$ [c64,c69]
$G_2(1352)$ [c23,c43]
$G_3(670)$ [c11,c16]
$G_{T_2}(1644)$ [c29,c63]
Feature extraction
HOG features pool
Joint-HOG features candidate pool
Final classifiers

図 5 Two-stage Real AdaBoost based on joint features

(2) SWT is more suitable for classification. Figure 4 (b) also shows the average gradient images of computed from 200 detection window images with or without applying SWT. In the average gradient image without applying SWT, the human silhouette and contours, especially the shoulders and legs are unclear. On the contrary, in the average gradient image with applying SWT, the human body contours are symmetric and easily recognized.

(3) The SWT contains the positions of the person in the world coordinate. These positions are identified and can be utilized to extract trajectories and analyze their behavior in the ISZOT.

2.2 Human detection using edge-based classifier

In conventional human detection methods, the symmetry and continuity of human shapes are difficult to classify. Normally, human shapes can be broadly divided into the following two characteristics:

(1) The Ω-shaped head and shoulder region and its continuation through the upper and lower body.

(2) The left-right symmetry of the head, shoulders, torso, legs and so on.

In this paper, we introduce a new object detection method [3] based on a two-stage Real AdaBoost [4] and joint features with which it is possible to automatically capture characteristics of type (1) and type (2). These joint features are made by using two-stage boosting to combine the HOG [5] features of two different regions. Figure 5 illustrates the creation of joint features and the final classifier structure. The creation and learning of joint features is performed using a two-stage Real AdaBoost algorithm. Here, we separately describe the first-stage Real AdaBoost that creates the joint features, and the second-stage Real AdaBoost that trains the final classifier.

Firstly, we create the joint features using a first-stage Real Adaboost. To create joint features, individual features are extracted from each of two different regions. Secondly, we construct the final classifier using a second-stage Real Adaboost. In the second-stage Real AdaBoost, the final classifier is constructed based on the input from the joint feature candidate pool created by the first-stage Real AdaBoost. In this way, it is possible to automatically select joint features that are useful for classification. Finally, all detection windows created by the SWT are classified using this final classifier.

2.3 Integration of classification results

In conventional human detection methods, integration of classification results in the image coordinate fails to distinguish human beings who are closed to each in the image view. Mean-shift [6] is widely used to integrate classification results into the final human detection result. However, the size of the kernel window varies with the camera setting and it is difficult properly adjust it in the image coordinate.

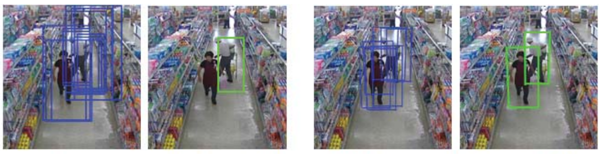
To overcome this problem, integration in the world coordinate is proposed. A Mean-shift in the world coordinate is implemented. In this method, the size of the kernel widow is adjusted easily based on the distance between human beings, which is robust to camera setting. For any candidate detected position of a person, A_j, the destination position A_{j+1} is calculated as:

$$\overrightarrow{A}_{j+1} = \frac{\sum_{i=1}^{n} \overrightarrow{A}_i g\left(\left\|\frac{\overrightarrow{A}_j-\overrightarrow{A}_i}{h}\right\|^2\right)}{\sum_{i=1}^{n} g\left(\left\|\frac{\overrightarrow{A}_j-\overrightarrow{A}_i}{h}\right\|^2\right)}, \overrightarrow{A}_i = (x_i, y_i, z_i) \tag{4}$$

where, $|| \;\; ||$ is the vector norm, and A is a position in the world coordinate.

Here, an isotropic Gaussian Kernel function $k(x)$ is selected. $g(x) = -k^{'}(x)$ denotes the derivative of this selected kernel profile, h denotes the size of the kernel window. A_j is shifted to the final result until A_{j+1} converges to zero. This process integrates similar results near a person efficiently. As shown in Figure 6, compared to the conventional Mean-shift in the image coordinate, the proposed approach has the following two features:

(1) It is easier to adjust the size of the kernel window h in the world coordinate: In conventional Mean-shift, h is difficult to perceive. It varies with different camera settings and the positions of people in the image. In order to obtain better integration results, h has to be adjusted properly. On the other hand, in the proposed Mean-shift in the world coordinate, h is adjusted as a constant value based on the distance between different people. This distance is easily estimated through the observed area and camera settings are irrelevant.



(a) Conventional Mean-shift (b) Proposed Mean-shift

図 6 Features of Mean-shift in the world coordinate

(2) Closed human figures are more properly identified. As show in Figure 6 (a), although two people are successfully classified, the conventional method still fails to identify them individually. On the contrary, as shown in Figure 6 (b), two people are identified properly. The proposed Mean-shift clustered classification results correctly identify each person's position in the world coordinate.

3 Trajectory Extracting

After human detection, a shopper's trajectory is extracted. In ISZOT, a simple but fast trajectory-extraction method is implemented. It is based on tracing a shopper's positions in the world coordinate, using Nearest Neighbor.

A trajectory P is extracted by sequentially searching the nearby area for the nearest shopper's positions. The nearest position, p_i, relative to the former position, p_{i-1}, is determined according to the Euclidean distance.

$$\begin{aligned} ||p_i - p_{i-1}|| &= \sqrt{(p_i - p_{i-1})^2} \\ &= \sqrt{(X_i - X_{i-1})^2 + (Y_i - Y_{i-1})^2} \end{aligned} \tag{5}$$

where, p_i, p_{i-1} is the shopper's standing position at time t_i and time t_{i-1}. The position is represented by$\{X, Y, Z\}^T$ in the world coordinate. Z is the ground that is always equal to 0. Then, the trajectory P can be expressed as:

$$\begin{aligned} P &= \{p_i\} \\ p_i &= [X_i, Y_i, Z_i, t_i]^T, i = 1, 2....N \end{aligned} \tag{6}$$

In ISZOT, a zone trajectory can also be defined and extracted based on shoppers' trajectories that contain detailed information of the shoppers' positions. This zone trajectory divides shoppers' positions into zones defined by different analysis purposes. For example, a zone can be defined based on individual areas and so on. There are two advantages to introducing zones into shoppers' trajectories:

(1) It can add semantical labels into shoppers' trajectories, and help in understanding their purchasing behaviors.

(2) It can compress shoppers' trajectories, which contain all positions in their histories.

Zone trajectory and its relative researches are introduced in [7].

4 Evaluation Experiments

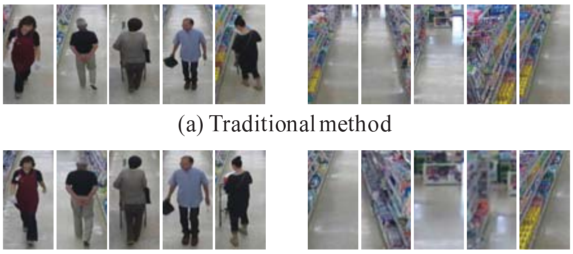
In this section, we evaluate our proposed system ISZOT with real surveillance data. These data were collected at the Yoichi store of the Consumers Co-operative in Sapporo.

4.1 Evaluation of human classification performance

In this paper, we proposed a new human classification method based on SWT and Joint-HOG. In order to confirm the effectiveness of this proposed method, we compare it with a traditional human classification method based on a Raster and Joint-HOG. Raster is a widely used scan method for human detection. Unlike SWT, it uses scalable rectangular windows to detect shoppers in the image coordinates.

Firstly, we created a training database for constructing human classifier. As shown in Figure 7, for the training database, 2394 positive data and 5000 negative data were collected. Figure 7 (a) shows the training samples created by the traditional method. Figure 7 (b) shows the training samples created by our proposed method. The positive data collected by each method correspond with each other. The negative data are random cuts from the same background image. After creating the training data, the classifier is constructed using the method introduced in Sec.2.2. Joint-HOG features are extracted for human classification.

For evaluation, 1273 positive data and 1856 negative data were collected separately using the traditional method and the proposed method. The evaluation result is shown in Figure 8. In this paper, a DET curve is introduced for evaluation. Compared to the traditional method, the proposed method classified more accurately. At a false positive rate of 1%, the miss rate was reduced from about 8.3% to 1.8%. This is because, due to the camera settings, the human shapes are distorted or tilted. In this situation, the traditional method failed to catch the human shapes we described in Sec.2 and failed to classify these features as human, such as legs, shoulders and so on. On the contrary, our proposed method has less distortion and is able to catch more human features to be classified.


(a) Traditional method

(b) Proposed method

図 7 Samples of positive and negative data

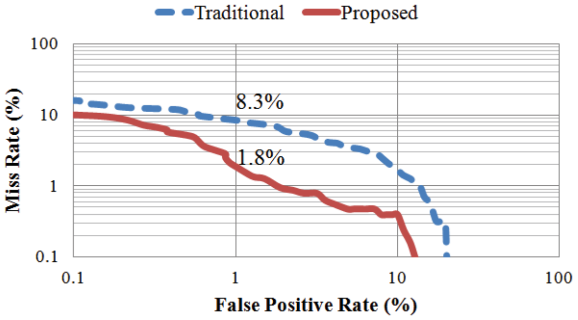


図 8 Result of DET curve

4.2 Evaluation of effectiveness of store layout using ISZOT

As a marketing experiment, we evaluated the effectiveness of the store layout by comparing the amount of shoppers' trajectories extracted from ISZOT.

In marketing, the store layout, or environment, is a key element to increase sales. A good store layout can get more shoppers into the store and increase the space productivity. Conventionally, the evaluation of store layout has always been carried out by people. They manually count the number of shoppers who enter an observed area and so on. It is very time consuming and the number of shoppers that can be evaluated is limited. Normally, it might take two staff people three days to collect information on 200 shoppers. However, in our experiments, instead of human observers, we evaluated a store layout using ISZOT.

Firstly, a main aisle was selected as the observed area. This aisle directly faced the store entry and is

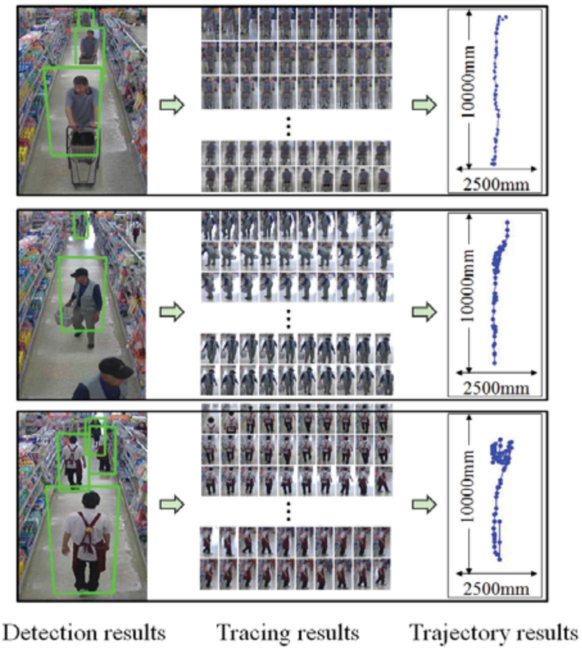


図 9　Example of processing results of ISZOT

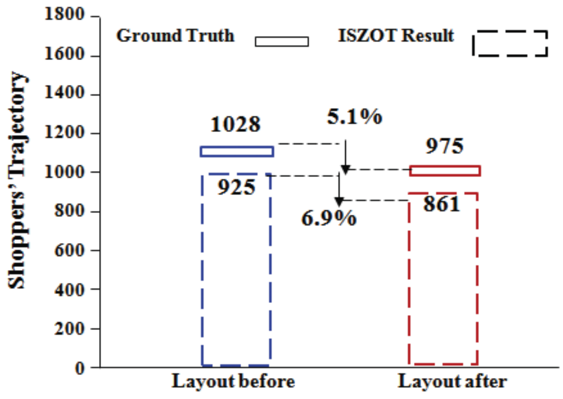


図 10　Counting results of shoppers' zone trajectories

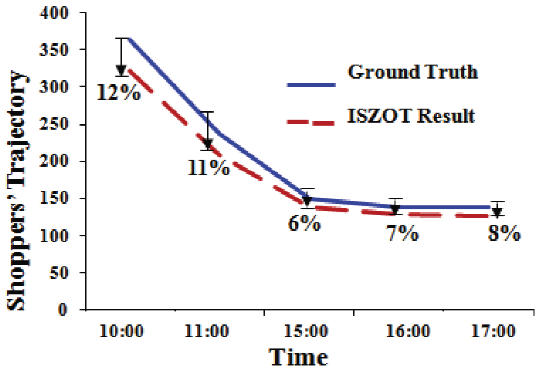


図 11　Extracting accuracy rate of shoppers's trajectories

expected to lead more shoppers to come inside other purchasing areas. The layout was rearranged specially for this evaluation experiment. Three weeks of surveillance data before and after this rearrangement were collected separately for comparison. Secondly, we selected data from the weekends (Saturdays and Sundays) and processed it using ISZOT for our experiment. These weekend days have more shoppers than weekdays and are difficult for human evaluators. Examples of processed results from ISZOT are shown in Figure 9. Shoppers with different shopping behaviors are detected and traced. Their detection results, tracing results and trajectories in this main aisle are extracted and presented. Among these shoppers, some of them walk through the main aisle and some of them stop for shopping and so on. In this experiment, about 1000 shoppers' trajectories before and after the layout rearrangement, in total about 2000 shoppers' trajectories were extracted. Finally, these extracted shoppers' trajectories were counted and compared to evaluate the layout rearrangement. Here, there might be several different parts of a shopper's trajectory. These parts were integrated and counted as one trajectory.

The evaluation results are shown in Figure 10. Surveillance data from six weekend days (three Saturdays and three Sundays) were processed both before the layout rearrangement and after. The solid line is the ground truth. The dashed line is the ISZOT result. We found that for the ground truth, after the layout rearrangement, the amount of shoppers' trajectories decreased by 5.1% compared to the former layout. For the ISZOT result, the amount decreased by 6.9%. The ISZOT result showed the same tendency as the ground truth did. Here, we analyzed the weekly fluctuation in the amount of shoppers' trajectories in the main aisle. In the results, any change under 13% is considered as normal fluctuation. Both decreases, 5.1% and 6.9%, are under 13%. From this point of view, we can conclude that there was no significant decrease after the layout rearrangement.

In this experiment, the extracting accuracy rate of shoppers' trajectories in ISZOT was evaluated hourly. The result is shown in Figure 11. The solid line is the ground truth. The dashed line is the ISZOT result. From 10a.m. to 12a.m., the main aisle was relatively crowded. About 350 shoppers passed this area. The accurate rate using ISZOT was around 88% (100%-12%). From 15p.m. to 17p.m., the number of shoppers decreased to 100. The accurate rate using ISZOT increased to approximately 93% (100%-7%). The average of the extracting accuracy rate was about 91%.

5 Conclusion

In this paper, we proposed a new surveillance and sensing system called ISZOT to analyze shoppers' purchasing behavior for marketing use. ISZOT takes advantage of image processing technologies. It is low cost and simple to use. It can analyze surveillance data collected anytime, anywhere, and including anybody in the stores. In this paper, some of the most essential technologies of ISZOT were introduced and explained. These essential technologies are: a new human classification method based on a SWT and Join-HOG was proposed, which is robust for surveillance camera settings; a new integration method based on Mean-shift in the world coordinate to detect shoppers' positions; and a simple method to extract shoppers' trajectories based on their positions. Experiments using surveillance data in Yoichi stores showed the effectiveness of these essential technologies. In this paper, ISZOT provides a new way for marketing analysis and is capable of more advanced analysis of purchasing behavior.

In future work, we are going to improve the analysis abilities of ISZOT. It will be interesting to combine the zone trajectories with POS data, since the POS data contains the actual shopping information at a particular time. This will help us to understand what shoppers intended to buy and what they did buy and so on. Also, in this paper, we only focused on one camera view. In the future, we would like to extend to multiple camera views. This will help us to understand a shopper's whole purchasing process in the stores. Finally, we expect that this system will become more precise and more powerful for marketing use.

6 Acknowledgment

The authors would like to specially thank Mr. K. Yoneda from the Consumers Co-operative in Sapporo for providing the experiment filed for this project and lots valuable advice for marketing analysis.

参考文献

[1] Y. Li, M. Ito, M. Miyoshi, H. Fujiyoshi, and S. Kaneko, Human Detection using Smart Window Transform and Edge-based Classifier, JSPE winter conference, 2011.

[2] R.Y.Tsai, A versatile Camera Calibration Technique for High-accuracy 3D Machine Vision Metrology using Off-the-shelf TV Camera and Lenses, IEEE Jounal of Robotics and Automation, Vol. RA-3, No. 4, pp.323-344, 1987.

[3] T. Mitsui, Y. Yamauchi and H. Fujiyoshi, Object Detection by Two-Stage Boosting with Joint Features, IEICE Transaction on Information and Systems, Vol. J92-D, No. 9, pp. 1591-1601, 2009.

[4] R. E. Schapire and Y. Singer, Improved Boosting Algorithms using Confidence-rated Predictions, Machine Learning, No.37, pp.297-336, 1999.

[5] N. Dalal and B. Triggs, Histograms of Oriented Gradients for Human Detection", IEEE Computer Vision and Pattern Recognition (CVPR), pp. 886-893, 2005.

[6] Y.Cheng, Mean shift, Mode Seeking and Clustering, IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.17, pp.790-799, 1995.

[7] T. Etchuya, H. Nara, S. Kaneko, Y. Li, M. Miyoshi, H. Fujiyoshi, and K. Watanabe, Extraction and Clustering of Zone-based Trajectories using Image Processing for Behavior Analysis, 2013 11th International Conference on Quality Control by Artificial Vision (QCAV), Fukuoka, Japan, 2013.