Spatial People Density Estimation from Multiple Viewpoints by Memory Based Regression

Yoshimune TABUCHI*, Tomokazu TAKAHASHI*, Daisuke DEGUCHI*, Ichiro IDE*, Hiroshi MURASE*,
Takayuki KUROZUMI† and Kunio KASHINO†
* Nagoya University
Furo-cho, Chikusa-ku, Nagoya-shi, Aichi, 464-8601, Japan
Email: tabuchiy@murase.m.is.nagoya-u.ac.jp, ttakahashi@gifu.shotoku.ac.jp,
ddeguchi@nagoya-u.jp, ide@is.nagoya-u.ac.jp, murase@is.nagoya-u.ac.jp
† NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation
3-1, Morinosato-Wakamiya, Atsugi-shi, Kanagawa, 243-0198, Japan
Email: kurozumi.takayuki@lab.ntt.co.jp, kashino.kunio@lab.ntt.co.jp

***Abstract*—Crowd analysis using cameras has attracted much attention for public safety and marketing. Among techniques of the crowd analysis, we focus on spatial people density estimation which estimates the number of people for each small area in a floor region. However, spatial people density cannot be estimated accurately for an area far from the camera because of the occlusion by people in a closer area. Therefore, we propose a method using a memory based regression method with images captured from cameras from multiple viewpoints. This method is realized by looking up a table that consists of correspondences between people density maps and crowd appearances. Since the crowd appearances include situations where various occlusions occur, an estimation robust to occlusion should be realized. In an experiment, we examined the effectiveness of the proposed method.**

I. Introduction

Crowd analysis has attracted much attention for public safety and marketing. Also, in recent years, it has become necessary to analyze a large volume of video data recording the activity of a crowd according to the widespread of surveillance cameras due to the increase of security awareness. Automatic analysis of such crowd videos is in demand because its manual analysis takes time and effort. Among techniques of the crowd analysis, we focus on spatial people density estimation using cameras. This is a technique which estimates the spatial distribution of the number of people from an input image as shown in Fig. 1. This will be beneficial to marketing because it is possible to obtain more detailed information such as the difference of the number of people between regions.

Various methods have been proposed to count people or track people for crowd analysis. As people counting methods using multiple cameras, a method using the detection results of a specific shape such as a human face [1], a regression based method using the relationship between image features and the number of people [2], and a method by counting people passing through a virtual gate defined manually on a screen [3] have been proposed. However, these methods do not consider how the people are spatially distributed in the camera's field of view because they aim to estimate just the number of people in it. On the other hand, for people tracking, a method using body parts [4], and methods by matching the same people between different cameras with no overlapped region in their fields of view [5][6] have been proposed. However, these methods are not aiming for people density estimation.

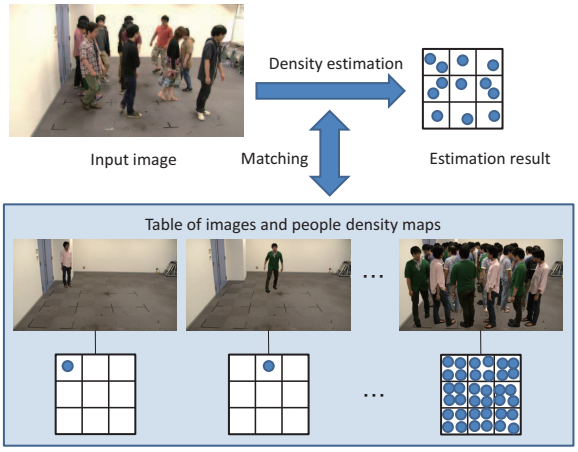
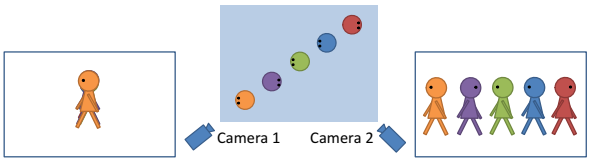


Fig. 1. Concept of spatial people density estimation from an image.



(a) Camera 1 (b) People distribution (c) Camera 2

Fig. 2. Example of a situation where people distribution cannot be perceived correctly by using only a single camera.

When estimating spatial people density from surveillance cameras, it will be difficult to do so accurately for an area far from the camera because of the occlusion by people in a closer area. To overcome this difficulty, we propose a method using a memory based regression method with

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.384

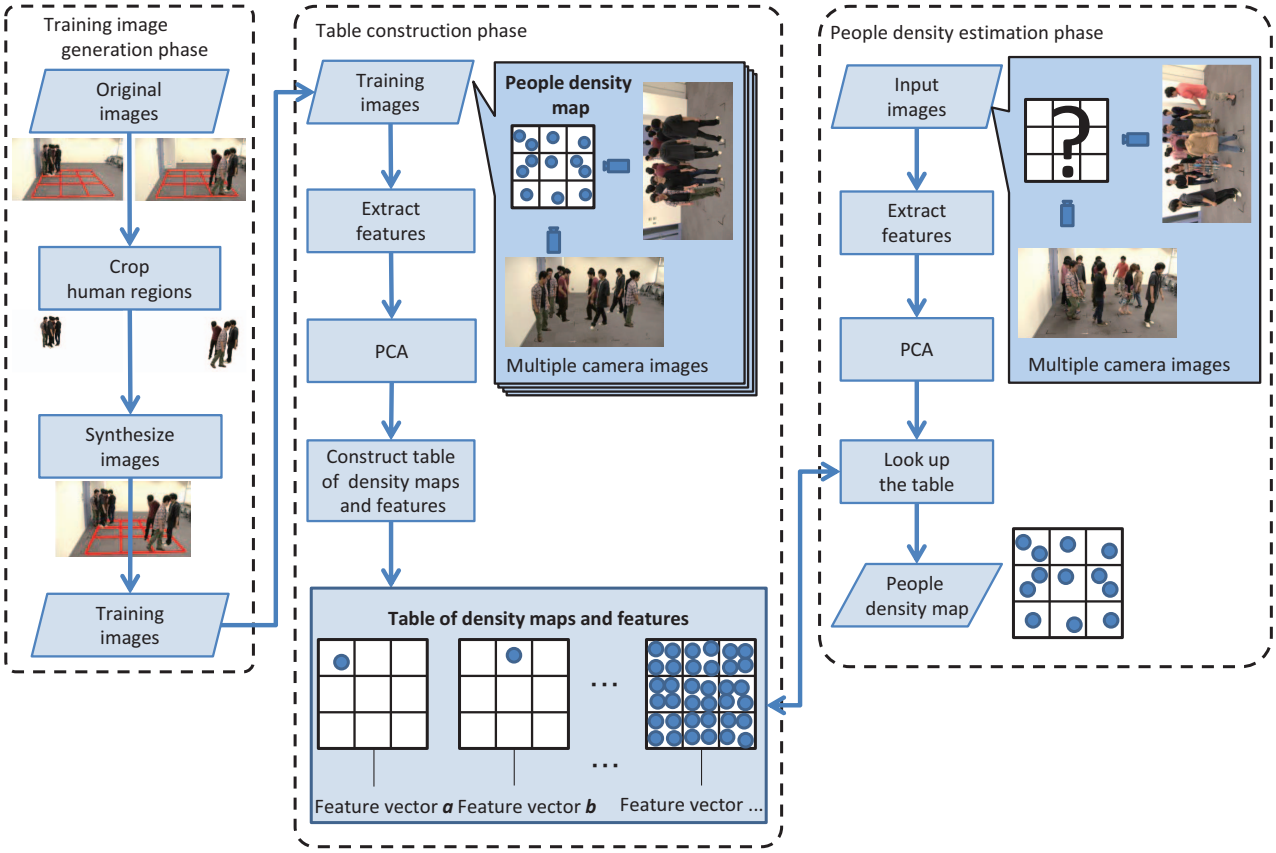


Fig. 4. Process flow of the proposed method.

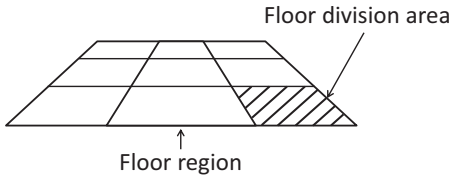


Fig. 3. Example of the division of a floor region.

images captured by cameras from multiple viewpoints. As shown in Fig. 1, the density estimation using the memory based regression method is realized by looking up a table that consists of correspondences between people density maps and crowd appearances. Since the crowd appearances include situations where various occlusions occur, an estimation robust to occlusion should be realized. In addition, we try to obtain more information for accurate estimation by using multiple cameras whose viewpoints are different and their fields of view overlap. Figure 2 shows an example of a situation where people distribution cannot be perceived correctly by using only a single camera. Here, if we captured an image of people that were distributed as illustrated in Fig. 2(b), from Camera 1, it will look like the image shown in Fig. 2(a). In this case, it is difficult to perceive the people distribution correctly due to occlusion by using only this single camera. On the other hand, if we captured the same people from Camera 2, we can perceive the people distribution correctly as the image shown in Fig. 2(c).

An approach which uses an overhead camera can also be employed in order to solve a similar problem easily. However, we did not take this option since we considered that circumstances that allow the installation of such cameras are limited.

In the following sections, we first describe the proposed spatial people density estimation method in Section II. We then explain the experiments and discuss the results in Section III. Finally, we conclude this paper in Section IV.

II. People Density Estimation by Memory Based Regression

The proposed method estimates the number of people for each of the small areas divided in a floor region. Hereafter, this area is called a "floor division area". In the following, we divide a floor region into 3×3 floor division areas as shown in Fig. 3.

Figure 4 shows the process flow of the proposed method. The proposed method consists of a training image generation phase, a table construction phase, and a people density

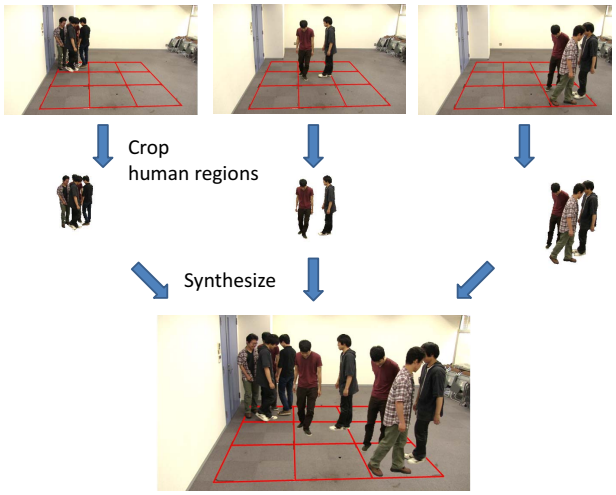


Fig. 5. Process flow for generating the training images.

Fig. 6. Example of a generated training image.

estimation phase. In the training image generation phase, training images are generated by an image synthesis strategy. In the table construction phase, image features are extracted from the training images. Then, they are registered to a table together with the corresponding people density map. In the people density estimation phase, the spatial people density is estimated by looking up the image features of input images in the table.

Although the proposed method can employ more than three cameras, in the following sections, we will explain the detailed procedure in the case of two cameras, for simplicity.

A. *Training image generation phase*

Training images are generated by the following process. Figure 5 shows the process flow of this phase.

1) Capture images of various scenes in advance where from 1 to 4 people exist in each floor division area.
2) Crop human regions manually from the captured images.
3) Synthesize the cropped human images to a background image to generate training images. The training images are generated for all people density maps that are used for the table construction in the next phase. Human regions cropped from each floor division area are overlapped from back to front according to a given people density map. Figure 6 shows an example of a generated training image. From this figure, we can see that the overlapping people between front and back floor division areas are naturally synthesized.
4) The process described above is performed for both of the two cameras.

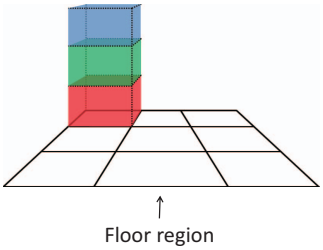


Fig. 7. Example of space division blocks.

B. *Table construction phase*

The table of density maps and features is constructed by the following process.

1) Extract image features from the space over each floor division area of the training images. Here, we divide the space over a floor division area into three blocks as shown in Fig. 7. We call the divided spaces as "space division block". This is because the properties of image features obtained from different parts of the body are different. From each space division block, we extract the following three image features: the number of pixels in the foreground, the number of edge pixels in the foreground, and the number of boundary pixels between the foreground and the background. Figure 8 shows an example of these image features. The number of the dimensions of a feature vector is 3 (image features) $\times$ 9 (floor division areas) $\times$ 3 (space division blocks) $\times$ 2 (cameras) $= 162$.
2) Reduce the dimensions of the extracted features. The PCA (Principal Component Analysis) method is used in order to select efficient features from the extracted features. Here, features whose eigenvalue is higher than 1 are used as principal components.
3) Construct a table of correspondences between the people density maps and the feature vectors after the dimension reduction.

C. *People density estimation phase*

People density is estimated by the following process.

1) Input images from two cameras that are configured at the same positions as in the table construction phase.

Fig. 8. Example of image features extracted from an input image.

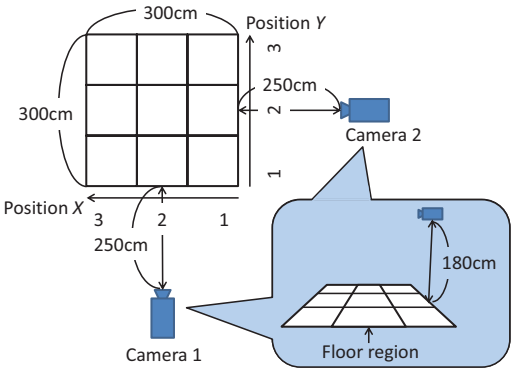


Fig. 9. Configuration of the two cameras and the floor division areas for the experiment.

2) Extract the image features from the input images and perform dimension reduction in the same manner as the table construction phase.
3) Find k-nearest neighbors of the input feature vector by comparing the input vector with vectors in the constructed table. Here, we used k-nearest neighbors instead of the nearest neighbor for the density estimation because different people density maps may have similar features.
4) Calculate the weighted sum of the people density maps corresponding to the k-nearest feature vectors as the estimation result. The weights are calculated based on the distances between feature vectors.

III. Experiments and Discussions

In order to investigate the effectiveness of the proposed method, we conducted an experiment on spatial people density estimation.

A. Experimental setup

Figure 9 shows the configuration of the two cameras and the floor division areas. In this experiment, we divided the floor region with a size of 300×300 cm^2 into 3×3 floor division areas. We used a foreground extraction method [7] and an edge detection method [8] for the image feature extraction. We set the parameter $k = 7$ for the k-nearest neighbors

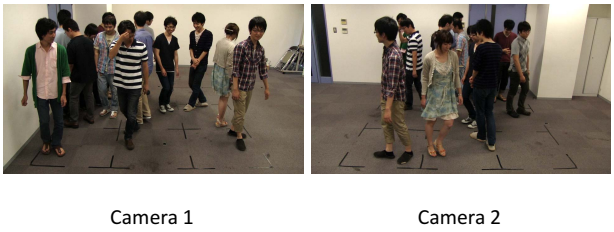
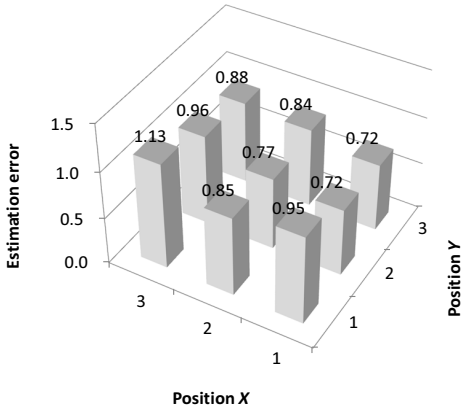
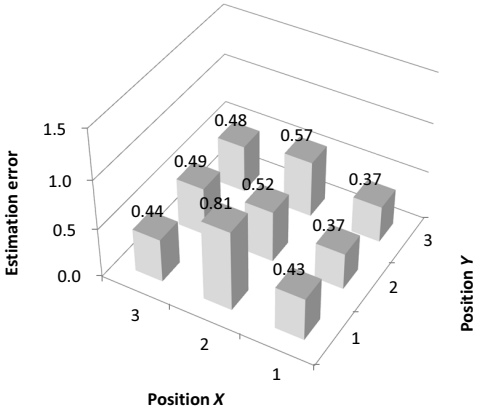


Fig. 10. Example of the test data.



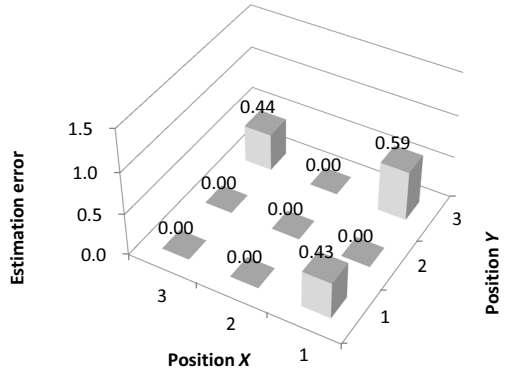
(a) The comparative method.



(b) The proposed method.

Fig. 11. Comparison of the estimation errors by floor division areas.

in the people density estimation phase. Mean absolute error of estimation results was used as an evaluation criteria. 184 images of from one to thirteen people which were captured simultaneously by two cameras were used as test data. The maximum number of people in a floor division area was four. Figure 10 shows an example of the test data. To generate the training images, we captured five images (one image each from zero to four people) for each of the nine floor

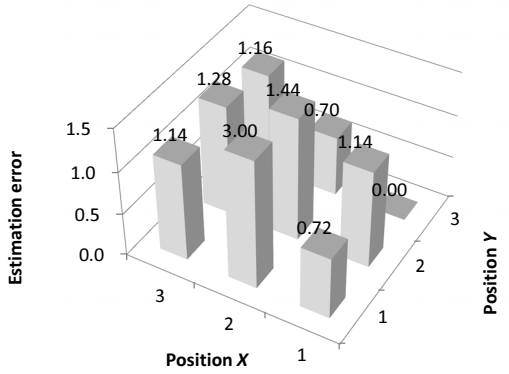


(a) Estimation errors by floor division areas.

(b) An input image.

(c) An image synthesized from the correct people density map.

Fig. 12. Example of a situation that the proposed method performed the estimation accurately.



(a) Estimation errors by floor division areas.

(b) An input image.

(c) An image synthesized from the correct people density map.

Fig. 13. Example of a situation that the proposed method failed the estimation largely.

division areas. As a result, the number of training images was $5^9 = 1,953,125$ in total.

We used a comparative method that estimates the number of people for each floor division area independently. This method learns the relationship between the extracted features and the number of people by a second order polynomial regression model in advance and performs the estimation using the learned model for each floor division area.

B. Results

Figure 11 compares the estimation errors by floor division areas between the proposed method and the comparative method. These graphs show that the accuracy of the proposed method was better than that of the comparative method for all floor division areas. This is because in the comparative method, it was difficult to estimate the people density accurately because of occlusion. On the other hand, in the proposed method, the table included situations where various

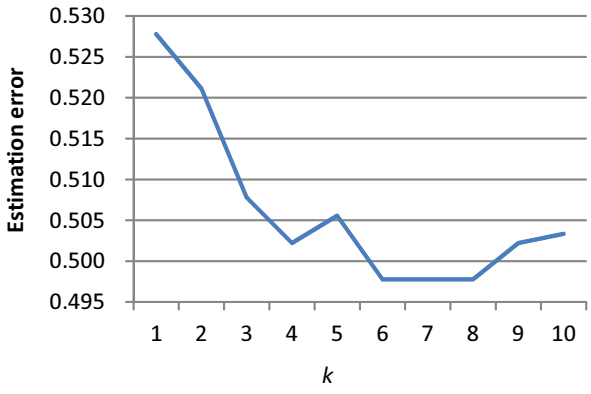


Fig. 14. The relationship between the estimation error and the parameter k.

occlusion occurred, which leaded to better accuracy. Therefore, we confirmed that the proposed method could estimate the density accurately even when an occlusion occurs, as intended.

Figure 12 shows an example of a situation when the proposed method performed the estimation accurately. Figure 12(a) shows the estimation errors by floor division areas, and Figs. 12(b) and 12(c) show the input image and the image generated from the correct people density map, respectively. We considered that the accuracy was high in this situation because the appearances of the input image and the generated image were similar. On the other hand, Fig. 13 shows an example of a situation when the proposed method failed the estimation largely. Figure 13(a) shows the estimation errors by floor division areas, and Figs. 13(b) and 13(c) show the input image and the image generated from the correct people density map, respectively. We can see that the appearance of these images are very different due to the difference of people positions and clothes even if they shared the same people density map. In the proposed method, the accuracy could degrade in such cases. This is because in this experiment, we used few original images for each situation in a floor division area to synthesize the training images. We consider that this problem could be dealt with by increasing the variation of the original images.

Figure 14 shows the relationship between the estimation error and the parameter k. The parameter k is the parameter of k-nearest neighbors in the people density estimation phase. This graph shows that the error becomes the minimum when $k = 7$. Therefore, we chose $k = 7$ as the parameter in the experiment.

IV. Conclusion

In this paper, we proposed a method for spatial people density estimation from multiple viewpoints by memory based regression. Specifically, the proposed method achieves people density estimation for each small area in a floor region by looking up the input image features in a table, which consists of correspondences between the people density maps and the image features that are extracted from synthesized training images.

In the experiment, the estimation accuracy of the proposed method was evaluated. From its result, we confirmed the effectiveness of the proposed method compared to the comparative method.

Future work will include the increase of the variations of the synthesized images and the use of other image features.

Acknowledgement

Parts of this research were supported by MEXT, Grant-in-Aid for Scientific Research. This work was developed based on the MIST library (http://mist.murase.m.is.nagoya-u.ac.jp/).

References

[1] M. Li, Z. Zhang, K. Huang, and T. Tan, "Estimating the number of people in crowded scenes by mid based foreground segmentation and head-shoulder detection," *Proceedings of 19th International Conference on Pattern Recognition*, no. WeAT2.1, pp. 1–4, December 2008.

[2] A. Chan, Z. Liang, and N. Vasconcelos, "Privacy preserving crowd monitoring: Counting people without people models or tracking," *Proceedings of 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1–7, June 2008.

[3] K. Terada, D. Yoshida, S. Oe, and J. Yamaguchi, "A method of counting the passing people by using the stereo images," *Proceedings of 1999 International Conference on Image Processing*, vol. 2, pp. 338–342, October 1999.

[4] B. Wu and R. Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, November 2007.

[5] O. Javed, Z. Rasheed, K. Shafique, and M. Shah, "Tracking across multiple cameras with disjoint views," *Proceedings of 9th International Conference on Computer Vision*, vol. 2, pp. 952–957, October 2003.

[6] O. Javed, K. Shafique, and M. Shah, "Appearance modeling for tracking in multiple non-overlapping cameras," *Proceedings of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 26–33, June 2005.

[7] S. Nagaya, T. Miyatake, T. Fujita, W. Ito, and H. Ueda, "Moving object detection by time-correlation-based background judgement method," *Transactions on IEICE D-II*, vol. J79-D-II, no. 4, pp. 568–576, April 1996, (in Japanese).

[8] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 6, pp. 679–698, November 1986.