Improving Accuracy of Bus Passenger Flows Estimation with Domain Adaptation

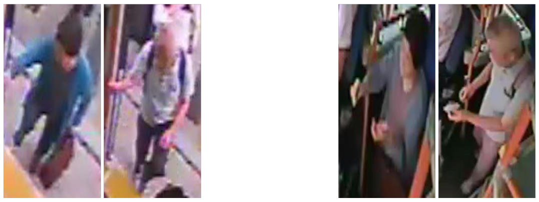
Aiwen Su

1 Introduction

BUScompanies survey the number of passengers get on and off the bus at each bus stop and use the results to formulate the schedules and routes of buses. One of the survey methods is to use sensors or records of IC cards. However, since sensors cannot identify individual passengers, it is not possible to record the distance traveled by each individual passenger. In addition, it is not possible to obtain data from passengers who do not use IC cards. To solve this problem, we propose to use computer vision technology to collect passenger data using surveillance cameras on the bus.

Recently, Suzuki et al proposed a method [1] to estimate the distance using passenger images captured by surveillance cameras. The flow of their method is: (1) Input the passenger images detected from the surveillance videos at the entrance and exit of the bus through the CNN(Convolutional Neural Network). (2) Extract the respective features. (3) Calculate the similarity between the two features. As a result, they obtained a 33.6% accuracy. However, there is a problem that their method does not consider the difference in the types of surveillance cameras on the bus. Particularly, as shown in Fig. 1, the entrance is RGB and the exit is near-infrared (IR), and the difference in color tone (domain difference) was not considered in Suzuki et al's method.

In this paper, we focus on aligning inputs from different domains to improve the accuracy of passenger re-identification on the bus. We attempted adversarial approach and domain-specific feature extractors and compare their performances. In addition, we propose a method using invariant features of passengers to reduce the effect of cross-domain. In the result, the proposed method improved the re-identification accuracy by 12.7%.



(a) Entrance(RGB) **(b)** Exit(NR)

Fig. 1: Different color tone in different cameras

2 Method

In this paper, we employ DANN(Domain Adversarial Neural Networks) [2] and AGW(Attention Generalized mean pooling with Weighted triplet loss) [3] to the conventional person identification method in order to extract features common to the entrance image (RGB) and exit image (IR) (i.e. common features). Furthermore, considering smaller color variations in the head part, we suppose that the head part is easier to extract common features than other parts and propose a method using the head features of passengers.

2.1 DANN

The architecture of DANN applied to person identification is shown in Fig. 2. First, a person image is input to ResNet, a feature extractor. Then, the person is classified using a label classifier and the cross entropy loss is calculated, then the parameters of the network are updated to extract similar features for persons with the same label. At the same time, the features extracted by ResNet are input to the domain classifier to improve the extraction ability of the network to extract common features between the two domains using the calculated loss. Furthermore, by using distance learning triplet loss, the network learns to extract similar features for the same person.

2.2 AGW

The architecture of AGW is shown in Fig. 3. Since AGW has a dedicated convolution layer for each domain, the

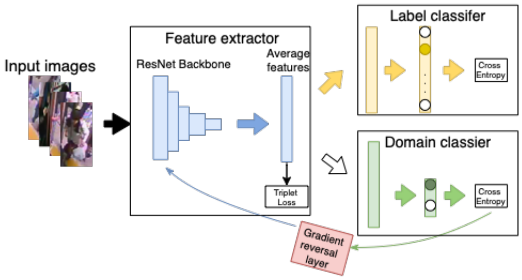


Fig. 2: Architecture of DANN

- *Aiwen Su is with the Department of Computer Science and Engineering, Waseda University, Tokyo, Japan.*
E-mail: baihe2333@gmail.com

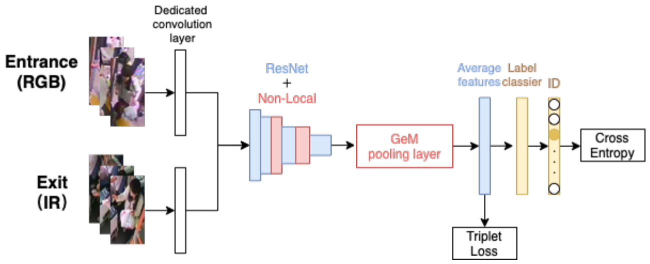


Fig. 3: Architecture of AGW

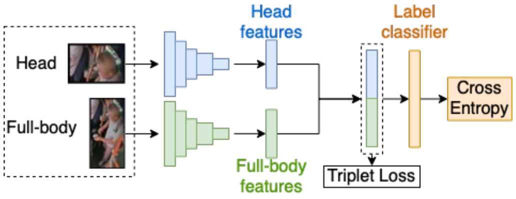


Fig. 4: Architecture of HeadReID

images captured by the RGB camera at the entrance and the IR camera at the exit are input separately for the first feature extraction. Next, they are input to ResNet50 and Non-Local for continuous feature extraction. They also pass through the GeM pooling layer and are classified using the all-coupling layer. In addition, the model is optimized with the cross entropy loss computed by the classification results and the triplet loss.

2.3 HeadReID

Since hair is not easily affected by color tones and the head has local features (e.g., headphones) that are easily distinguished, we propose a method (HeadReID) using the head features of person. The architecture of HeadReID is shown in Fig. 4. First, the head part is extracted from the input full-body image. Then, the full-body image and the head image are input to separate networks, and the resulting features are concatenated and input to the classifier. This makes it possible to consider the difference between global and local features.

3 Experiment

In this experiment, to demonstrate the effectiveness of domain adaptation methods, we compared the person identification accuracy of three methods, DANN, AGW, and HeadReID, with that of Suzuki et al. We also compared the performance of the three modules in AGW.

We used the video images taken inside the same bus for three days (16 sections in total). The experimental data are the images of people who appear within the range of the entrance and exit. The person images were divided into training data and test data. Specifically, one segment of the images of the daytime, nighttime, and congestion time was used as test data, and the remaining segments were used as training data. The learning rate was set to 0.0003. The accuracy rate was used as the evaluation method. Accuracy rate = (number of correct answers) / (number of query images).

3.1 Results and discussion

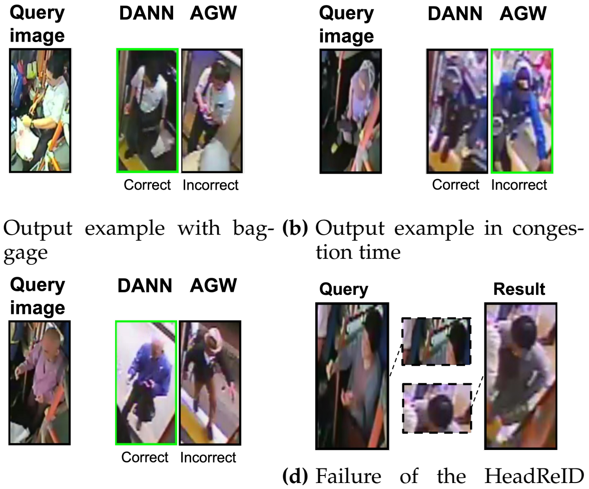
Table 1 shows the experimental results. The use of the domain adaptation method improves the percentage of correct responses in the daytime and nighttime compared to the method used by Suzuki et al. The following sections discuss the results for each method (DANN, AGW, and HeadReID).

TABLE 1: Comparison of person identification accuracy

Method	daytime	nighttime	congestion time
Suzuki et al [1]	75.0%	72.2%	52.5%
DANN [2]	83.3%	74.2%	47.5%
AGW [3]	**91.7%**	**80.0%**	**65.8%**
HeadReID	71.3%	68.5%	49.0%

TABLE 2: Comparison of accuracy rates for each module in AGW

module	accuracy rate
Dedicated convolutional layer	**79.4%**
NonLocal	54.8%
GeM	50.2%



(a) Output example with baggage

(b) Output example in congestion time

(c) Output example with hat

(d) Failure of the HeadReID method

Fig. 5: Architecture of three methods.

DANN. DANN showed a high accuracy rate during the daytime and nighttime, but the rate decreased during congestion. This is because as shown in Fig. 5a and 5b, DANN focuses on extracting common features from two domains, such as structural features of a person, and is not suitable for cases where persons overlap. It was also found that when there are no personal belongings such as a bag, recognition becomes difficult because features except full-body features cannot be extracted.

AGW. AGW had the highest accuracy rate in all time segments. The comparison of accuracy rates for each module shown in Table 2 indicates that the dedicated convolutional layer is the most effective. The separation of the first layer of ResNet may help reduce the weight of domain-specific features (i.e., color tones), which may lead to the extraction of common features. However, as shown in Fig. 5c, we could not focus on local features such as headphones and hats, which lead to errors.

HeadReID. This method did not achieve the same high accuracy as the other two methods. There were two possible causes: (1) The head image was improperly cropped; as shown in Fig. 5d, the head region was cropped from a fixed area, and part of the passenger's clothing was also included, which may affect the extraction of head features. (2) The low resolution made it difficult to extract the head features.

4 Conclusion

In this study, we conducted experiments to compare the accuracy of person identification among the three domain adaptation methods. AGW improved the accuracy rate by 12.7% in daytime, nighttime, and congestion time compared to the method of Suzuki et al. In addition, we introduced HeadReID to focus on local features that could not be focused on by AGW. However, the incorrect cropping of the head image resulted in a lower accuracy rate. In the future, we aim to improve the accuracy of head image cropping by using a head detector and to focus on local features.

References

[1] T. Suzuki and Y. Shimada and Y. Taniguchi, ”Human Pose Estimation and Motion Analysis for Estimating Bus Passenger Flows”, In IEVC, 4B-3, 2019.

[2] Y. Ganin and V. Lempitsky. Unsupervised Domain Adaptation by Backpropagation. In ICML, Vol. 37, pp. 1180-1189, 2015.

[3] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi. Deep Learn-ing for Person Re-identification: A Survey and Outlook. TPAMI, pp. 1-1, 2021. doi:10.1109/TPAMI.2021.3054775.