Unsupervised Domain Adaptation Learning Algorithm for RGB-D Stairway Recognition

Jing WANG[1], Kuangen ZHANG[1, 2]

(1. *Department of Mechanical Engineering, The University of British Columbia, Vancouver V6T1Z4*;
2. *Department of Mechanical and Energy Engineering, Southern University of Science and Technology, Shenzhen* 518055)

Abstract: Detection and recognition of a stairway as upstairs, downstairs and negative (e.g., ladder, level ground) are the fundamentals of assisting the visually impaired to travel independently in unfamiliar environments. Previous studies have focused on using massive amounts of RGB-D scene data to train traditional machine learning (ML)-based models to detect and recognize stationary stairway and escalator stairway separately. Nevertheless, none of them consider jointly training these two similar but different datasets to achieve better performance. This paper applies an adversarial learning algorithm on the indicated unsupervised domain adaptation scenario to transfer knowledge learned from the labeled RGB-D escalator stairway dataset to the unlabeled RGB-D stationary dataset. By utilizing the developed method, a feedforward convolutional neural network (CNN)-based feature extractor with five convolution layers can achieve 100% classification accuracy on testing the labeled escalator stairway data distributions and 80.6% classification accuracy on testing the unlabeled stationary data distributions. The success of the developed approach is demonstrated for classifying stairway on these two domains with a limited amount of data. To further demonstrate the effectiveness of the proposed method, the same CNN model is evaluated without domain adaptation and the results are compared with those of the presented architecture.

Key words: Domain Adaptation, convolutional Neural Network, Deep Learning, RGB-D Scene Data, Stairway Classification, Visually Impaired.

1 Introduction

Approximately 285 million people around the world were visually impaired, of whom 39 million were blind, based on the 2010 World Health Organization survey [1]. Although some mobility assistant systems, which are based on converting sonar information into audible signals, have been developed to facilitate navigation, obstacle detection, and wayfinding tasks [2], the visually impaired still face challenges to actively interact with dynamic surrounding environments, such as stairways. Stairways widely exist in both indoor and outdoor environments. Thus, the potential risk of falling from stairs, especially downstairs, can be fatal to the visually impaired.

To effectively detect stairways, monocular cameras, stereo cameras, and laser scanning devices (e.g., LiDAR) have been used. However, the existing approaches can only provide user-selected information instead of autonomously taking into account the obstacles presented in motion [8]. RGB-D cameras, which can simultaneously capture both visual features and depth information of the environment, are widely used to recognize indoor and outdoor environments [3-4]. Many researchers have focused on combining visual features with depth information for robust image representation [5-9]. He et al.[9] directly combined RGB channels with the corresponding depth information at the early stage to generate four-channel RGB-D images (early fusion). By feeding their RGB-D images in a feedforward CNN, they obtained a higher validation accuracy compared with that of feeding the original RGB images into the same CNN architecture. This research implies that depth channel has a better feature representation than the R, G, B channels. However, their training and testing samples come from the same data distribution. Thus, if their model is utilized for testing another dataset, its performance will degrade significantly [10]. Munoz et al. [8] and Wang et al. [7] constructed the one-dimensional depth vector as parallel lines, which contains the distance from a camera and the orientation of each stair, to train a support vector machine (SVM) classifier and classify upstairs, downstairs, and level ground. They both used edge detection to gain edge maps from RGB images and extracted one-dimensional depth vectors with the distance and orientation information from edge maps by applying Hough transform.

Nevertheless, the performance of edge detection algorithms and Hough transform highly depends on the level of chosen thresholds. In this manner, they have carried out numerous preprocessing tasks to make their datasets adapt to their model. Thus, their model will poorly perform when it is tested on real-case scenarios, whose data distribution can be significantly different from that of their dataset.

Domain adaptation (DA) refers to a situation where one aims at learning a discriminative classifier from samples drawn from the source and applying this classifier to a different but related target data distribution. DA methods can learn a mapping between the source domain and the target domain when the target samples are either fully unlabeled (unsupervised domain adaptation) or partially labeled (semi-supervised domain adaptation). The present paper proposes an adversarial learning method for a domain adaptation scenario in the presence of a shift (as shown in Fig. 1) between two data distributions. The task is formulated as an unsupervised domain adaptation (UDA) problem, in which escalator stairway visual features and depth features are labeled in the source domain while having the unlabeled stationary stairway visual features and depth features in the target domain.

To apply feed-forward networks to the stationary stairway (target) domain without being hindered by the shift between the two domains, it is required to embed domain adaptation into the process of learning representation so that the classification decisions are made based on features that are both discriminative and invariant to the change of domains. Inspired by domain adversarial neural networks (DANN) [11], a feature extractor, label classifier, and domain classifier are embedded into a composed deep feedforward network to classify the stairway images across the two domains. Additionally, the unsupervised domain adaptation is achieved by adding a domain classifier connected to the feature extractor via a gradient reversal layer. The gradient reversal layer ensures that the feature extractor only extracts domain-invariant features by multiplying the gradient by negative one during the backpropagation.

The main contributions of the paper are as follows: 1) the possibility of using adversarial learning methods to tackle unsupervised domain adaptation scenario from the escalator to stationary stairway is verified; 2) an adversarial learning method is embedded into a deep feedforward convolutional neural network for unsupervised domain adaptation.

The remainder of the paper is organized as follows: Section 2 discusses the proposed framework and the experimental methods of the present work. Section 3 presents the experimental results. Section 4 discusses the limitations of the work. Section 5 concludes the paper and indicates the planned future work.

2 Adversarial Learning for Unsupervised Domain Adaptation

For classification tasks, the input space is X and the set of L possible labels is Y. In the setting of unsupervised domain adaptation, n labeled examples are sampled from the escalator distribution $P(X^s, Y^s)$ to form the source domain Ds = $\{(X_i^s, Y_i^s)\}_{i=1}^{n}$. Similarly, m unlabeled examples are sampled from the stationary data distribution $Q(X^t, Y^t)$, which is different but similar to the escalator data distribution, to form the target domain Dt = $\{(X_j^t)\}_{j=1}^{m}$.

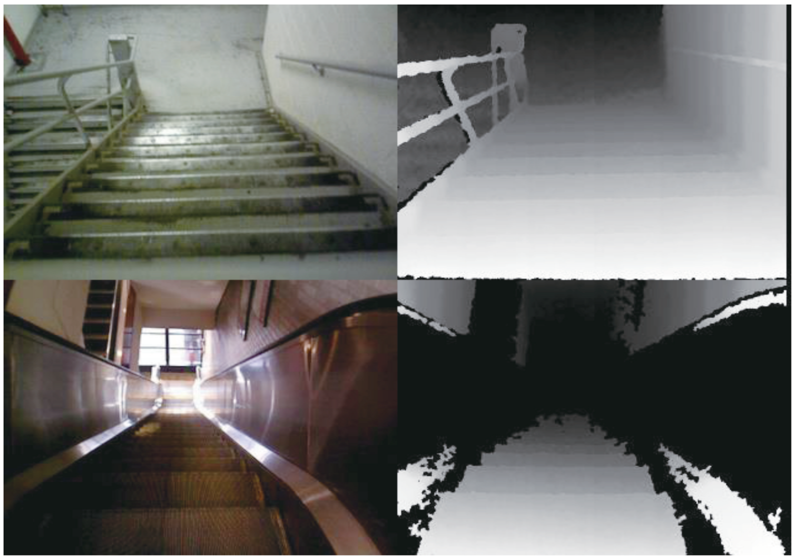

Fig. 1 Examples of the RGB image and the depth image of the downstairs stationary stairway dataset are shown on the top-left and top-right corners, respectively. Examples of the RGB image and the depth image of the downstairs escalator stairway dataset are shown on the down-left and down-right corners, respectively. An obvious shift between the two data distributions can be observed.

The work of Ben-David et al. [12] shows that the target risk R_T is upper bounded by three items: 1) the source risk R_S; 2) the divergence of the hypothesis space $d_H(S,T)$; 3) a constant complexity term that depends on the VC-dimension d and the size of the samples S and T, which is denoted by m.

Theorem 1 [12] Let H be a hypothesis space of VC-dimension d. With a probability of at least 1-δ (over the choice of samples $S \sim D_S$ and $T \sim D_T$), for every h ∈ H:

$$R_T \leq R_S + d_H(S,T) + \frac{4}{m}\sqrt{\left(dlog\frac{2m}{d} + log\frac{4}{\delta}\right)} + 4\sqrt{\frac{dlog(2m) + log\frac{4}{\delta}}{m}}$$

The theorem indicates that the learning algorithm should minimize a trade-off between the source risk and the divergence $d_H(S,T)$ to minimize the empirical target risk. This paper aims to learn a feature extractor $f = G_f(x,\theta_f)$ that extracts domain-invariant features, a label predictor $y = G_y(f,\theta_y)$ that computes label predictions, and a domain classifier $G_d(f,\theta_d)$ that computes domain predictions and reduces the shift between two joint distributions. So, the target risk $R_T = \frac{1}{m}\sum_{i=1}^{m} Pr_{(x,y)\sim Q}[G_y(G_f(x_i)) \neq y_i]$ can be minimized by jointly minimizing the source risk and the distribution difference.

2.1 Domain Adversarial Neural Networks

Domain adversarial neural networks explicitly implement Theorem 1 into a neural network classifier that can extract transferable features to reduce the distribution shift between the two domains. To extract as many transferable features as possible, a deep feedforward convolutional neural network is designed as the feature extractor $G_f(x,\theta_f)$, with parameters θ_f, for extracting discriminative features of images [14]. This paper uses cross-entropy loss with softmax function to define the discrepancy between predictions and the original input distributions. To ensure that the extracted features are domain-invariant, the parameter θ_f is jointly learned by maximizing the loss of domain predictor $G_d(f,\theta_d)$ and minimizing the loss of label predictor $G_y(f,\theta_y)$. The label prediction loss and the domain prediction loss,

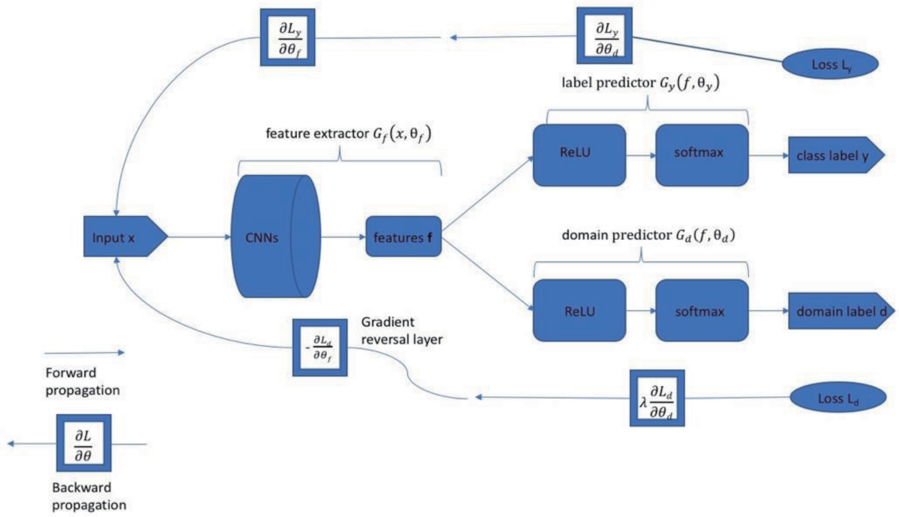

Fig. 2 Domain adversarial neural network includes a standard feed-forward architecture formed by a deep feature extractor and a deep label predictor. The shift in two joint distributions is reduced by adding a domain predictor connected to the feature extractor via a gradient reversal layer. The gradient reversal layer multiplies the gradient by negative one during the backpropagation training, which makes the feature distributions over the two domains similar.

respectively, are denoted by:

$$L_y(\theta_f,\theta_y) = L_y(G_y(G_f(x_i,\theta_f),\theta_y),y_i) \quad (1)$$

$$L_d(\theta_f,\theta_d) = L_d(G_d(G_f(x_i,\theta_f),\theta_d),d_i) \quad (2)$$

Thus, the domain adaptation problem becomes optimization of the objective function

$$R(\theta_f,\theta_y,\theta_d) = \frac{1}{n_s}\sum_{x_i\in D_S} L_y(\theta_f,\theta_y) - \frac{\lambda}{n_t+n_s}\sum_{x_i\in(D_S\cup D_T)} L_d(\theta_f,\theta_d) \quad (3)$$

where n_s and n_t are the number of examples in the source domain and target domain, respectively; and the loss of the domain classifier is weight by λ. By optimizing the objective function (3), the parameters θ_f, θ_y, θ_d will deliver a saddle point ($\hat{\theta}_f$, $\hat{\theta}_y$, $\hat{\theta}_d$)

$$(\hat{\theta}_f,\hat{\theta}_y) = \arg\min_{\theta_f,\theta_y} R(\theta_f,\theta_y,\hat{\theta}_d) , \quad (4)$$

$$(\hat{\theta}_d) = \arg\max_{\theta_d} R(\hat{\theta}_f,\hat{\theta}_y,\theta_d) . \quad (5)$$

by the following gradient updates:

$$\theta_y - \mu\left(\frac{\partial L_y}{\partial \theta_y}\right) \rightarrow \theta_y \quad (7)$$

$$\theta_d - \mu\lambda\left(\frac{\partial L_d}{\partial \theta_d}\right) \rightarrow \theta_d \quad (8)$$

where μ is the learning rate.

2.2 CNN Architecture for Feature Extractor

To perform domain adversarial training, a deep convolutional neural network is constructed for feature extractor $G_f(x,\theta_f)$. Ganin et al.[11] experimented their domain adversarial neural network on MNIST and MNIST-M datasets. The two data distributions have much fewer features and less noise compared with the real-world stairway datasets used in the present paper. Thus, the present feature extractor should be deep enough, which is able to extract discriminative features of stairway datasets.

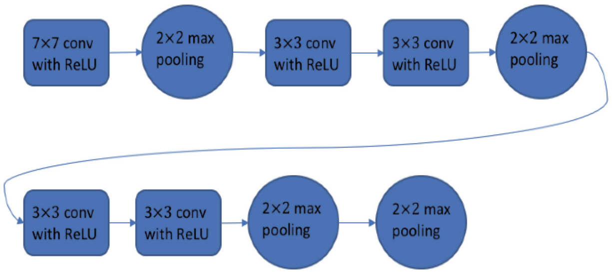

Fig. 3 The proposed deep convolutional neural network architecture for the feature extractor. The input of the feedforward architecture consists of 80 × 80 × 4 RGB-D images. The output is a 1200-dimensional (5 × 5 × 48) vector which contains discriminative features.

The feature extractor $G_f(x,\theta_f)$ incorporates five convolution layers and one fully-connected layer, which gives 1200-dimensional feature vector as an output. The first convolution layer applies 32 7 × 7 filters (extracting 7 × 7-pixel sub-regions) with ReLU activation function and 2 × 2 max pooling. The second convolution layer applies 32 3 × 3 filters with ReLU activation function. The third convolution layer applies 48 3 × 3 filters with ReLU activation function and 2 × 2 max pooling. The fourth convolution layer applies 48 3 × 3 filters with ReLU activation function. The fifth convolution layer applies 48 3 × 3 filters, with ReLU activation function, followed by two 2 × 2 max-pooling layers. During each training batch, the pairwise cosine similarities between 1200-dimensional features are calculated and back-propagated as the loss for all pairs within the batch.

A label predictor $G_y(f,\theta_y)$ and a domain predictor $G_d(f,\theta_d)$ are attached to the 1200-dimensional bottleneck of the fully-connected layer of the feature extractor in parallel (as shown in Fig. 2). The label predictor consists of three fully connected layers (1200 → 100 → 100 → 3). Similarly, the domain predictor has two fully-connected layers (1200 → 100 → 2). θ_f is learned by jointly training the label predictor and the domain predictor. Thus, domain-invariant features can be effectively extracted by the present feature extractor after the training convergence.

2.3 RBG-D Data Generation

As discussed above, depth information has a better feature representation than R, G, B channels. So, an early fusion is applied on the RGB and depth channels to construct the four-channel input. To add depth information, the original three-channel RGB images are extended to four channels and the corresponding depth information is copied to the fourth channel.

3 Results

1,105 escalator RGB-D image pairs and 2,157 stationary RGB-D image pairs are selected from RGB-D Stairway Detection Dataset,[8] which consists of three classes (downstairs, upstairs, and negative cases). To combine RGB images with its corresponding depth information, all RGB images are resized to 72 × 72 × 4 and the corresponding depth information is copied to their fourth channel. The escalator stairway dataset (source domain) is split into a training set (80%) and a testing set (20%). The stationary stairway dataset (target domain) is split into a training set (40%) and a testing set (60%). Each data sample is a 72 × 72 × 4 matrix. The learning rate is adjusted during the stochastic gradient descent by $\mu = \dfrac{\mu_0}{(1+\alpha \cdot p)^{\beta}}$. The learning rate could alternatively be adjusted by cosine decay $\mu = \dfrac{1}{2}\left(1+\cos\dfrac{t\pi}{T}\right)\mu_0$ at batch t, where T is the total number of batches [17]. During the training process, the initial learning rate μ_0, momentum p, and the batch size are set as 0.0005, 0.45, and 256, respectively. To promote convergence and low error on the source domain training, the following values are used: $\alpha = 10$ and $\beta = 0.75$. The domain adaptation parameter λ is initiated at 0. During the training process, λ is updated through $\lambda = \dfrac{2}{1+exp(-10p)}-1$.

3.1 Different Stairways

Three types of the stairway are shown in Fig. 4. The difference between various types of stairway can be easily distinguished. By comparing the stairway RGB image with its corresponding depth information, it is argued that the depth information has better feature representation and less noise than the RGB information. Thus, more discriminative features can be extracted from the depth channel rather than from the R, G, B channels.

3.2 Results without Domain Adaptation

When relatively large amounts of training data are available, CNN can learn more discriminative features than any other existed methods do [15]. To show the advantage of using CNN to extract discriminative features, the proposed CNN architecture is trained and tested on the escalator stairway dataset (source domain).

Table 1 Experimental Results of the Escalator Stairway Recognition with the CNN architecture.

Prediction / Ground Truth	Upstairs	Downstairs	Negative	Accuracy
Upstairs	79	1	0	98.75%
Downstairs	0	79	1	98.75%
Negative	0	0	63	100.0%

Table 2 Experimental Results of the Stationary Stairway Recognition with the CNN architecture only trained on the Escalator Dataset.

Prediction/ Ground Truth	Upstairs	Downstairs	Negative	Accuracy
Upstairs	244	130	95	52.03%
Downstairs	82	281	74	64.30%
Negative	77	63	248	63.92%

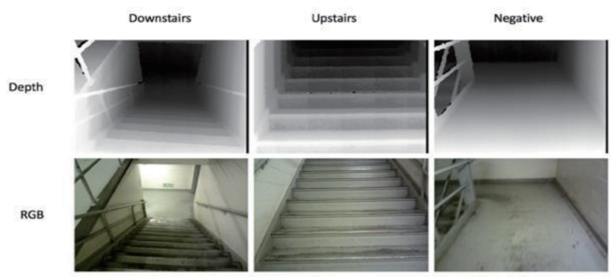

Fig. 4 Different stationary stairways. The first row shows depth images; the second row shows the corresponding RGB images.

The results (as shown in Table. 1) indicate that the proposed CNN can achieve an overall test accuracy of 99.28% on the source domain. After comparing the performance of the proposed CNN with that of the Munoz et al.' s model (92.7%), one can conclude that it is advantageous to use a deep feed-forward CNN to extract discriminative features for the four-channel RGB-D images. The results (overall 59.72%) shown in Table. 2 indicate a significant degrading of performance when the model trained only on the source is tested on the target.

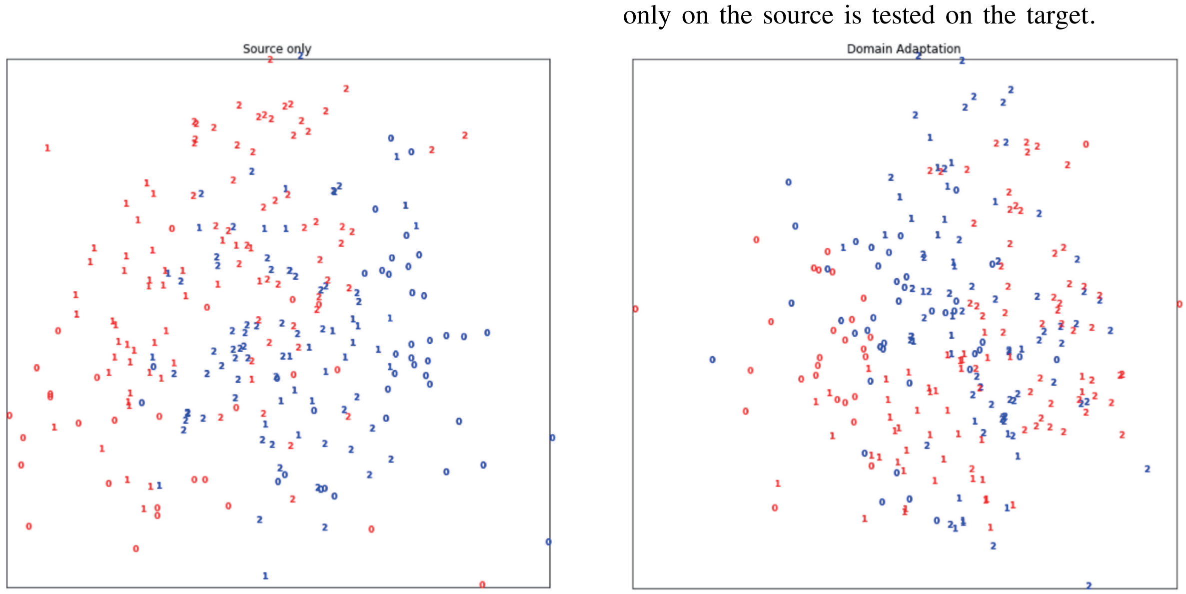

Fig. 5 The effect of domain adaptation on the distribution of extracted features. The left figure illustrates the feature distributions before domain adaptation. The right figure illustrates the adapted feature distributions. Blue points correspond to the escalator stairway data distributions (source domain). Red points correspond to the stationary stairway data distributions (target domain). The number 0 represents the downstairs case; the number 1 represents the upstairs case; the number 2 represents the negative case.

3.3 Results with Domain Adaptation

The proposed system incorporates the DANN regularizer with the proposed feed-forward CNN architecture. To mitigate the shift between two data distributions, the labeled escalator stairway data is combined with the unlabeled stationary stairway data to train the proposed system. The present model achieves an average accuracy at 100% for the escalator (source domain) stairway recognition. This result indicates that the domain adaptation regularizer can improve the performance of the classifier by adapting the escalator scenario to the stationary scenario.

After the training convergence, the present model is tested on the stationary stairway dataset (target domain). The results shown in Table 4 indicate that the present model successfully mitigates the effect of the data distribution shift with an average test accuracy at 80.6% for the stationary stairway recognition.

Table 3 Experimental Results of the Escalator Stairway Recognition with domain adaptation.

Prediction/ Ground Truth	Upstairs	Downstairs	Negative	Accuracy
Upstairs	80	0	0	100.0%
Downstairs	0	80	0	100.0%
Negative	0	0	63	100.0%

Table 4 Experimental Results of the Stationary Stairway Recognition with domain adaptation.

Prediction / Ground Truth	Upstairs	Downstairs	Negative	Accuracy
Upstairs	381	56	32	81.24%
Downstairs	47	339	51	77.57%
Negative	25	40	323	83.25%

3.4 Visualizations of Domain Adaptation

To visualize different feature distributions between the source and the target, t-SNE projection[16] is applied on the last hidden layer of the label predictor $G_y(f, \theta_y)$. As shown in Fig. 5, the domain adaptation regularizer makes two different feature distributions get closer. This implies that the feature extractor $G_f(x, \theta_f)$ has been successfully confused by jointly training the label predictor $G_y(f, \theta_y)$ and the domain predictor $G_d(f, \theta_d)$.

The overlap between the different distributions in Fig. 5 indicates the success of domain adaptation. Moreover, it is observed that the overlap also corresponds to the classification accuracy for the target domain, i.e., more overlap in t-SNE projection results in higher classification accuracy for the target.

3.5 Computational Environment

The training and testing of the present model were implemented on a computer with an AMD Ryzen 7 2700X Eight-Core Processor (3.7 GHz), a 16 GB DDR3, and an NVIDIA GeForce GTX 1060 graphics card. This computational environment allows the model to be trained on 256-sized batches. The labeled samples from the source constitutes half of each batch. The rest of the batch constitutes the unlabeled samples from the target.

4 Discussion

This paper incorporated domain adaptation methods into stairway recognitions. To extract more transferable features from two different data distributions, a deep feedforward convolutional neural network was designed to learn more discriminative features of the RGB-D stairway dataset. The results indicated that, without using domain adaptation techniques, the proposed CNN architecture could achieve 99.28% test accuracy on source domain data distributions (the escalator stairway dataset). Furthermore, the present RGB-D stairway dataset has much less labeled data compared with the Munoz et al.'s stairway dataset. However, the proposed CNN architecture still outperforms their model.

This work also indicates that domain adaptation methods will contribute to a better training convergence. By jointly training the label predictor with data from two different distributions, the model will outperform the proposed CNN architecture, and converge with fewer training epochs.

Although the proposed CNN architecture with domain adaptation regularizer can classify the stairway at high accuracy and mitigate the effect of data distribution shift, it is acknowledged that there are still some limitations. First, one should expand the categories of the stairway, such as obstacle, ramp, and wall, to enhance the environmental adaptability of the proposed model in more complex environments. Besides, the present model adopted only two different stairway data distributions. More stairway data distributions should be considered in the future work so that the model will be more robust. Finally, the proposed method has only been evaluated in the offline analysis.

5 Conclusion

This paper present a deep architecture deep architecture for domain adversarial neural networks to transfer the knowledge learned from the labeled escalator stairway data distributions to the unlabeled stationary stairway data distributions. Unlike the previous stairway recognition methods, the accuracy of the proposed model did not rely on a large amount of labeled data. The results demonstrated that the developed model would achieve better performance with much less labeled data compared with other methods. The developed model successfully mitigated the degradation of performance caused by the shift between data distributions. Moreover, the present work indicated that RGB-D images had better feature representations than RGB images.

References

[1] *Universal eye health: a global action plan* 2014-2019. Geneva: World Health Organization, 2013.

[2] M. Bousbia-Salah, A. Redjati, M. Fezari, and M. Bettayeb, "An Ultrasonic Navigation System for Blind People," in 2007 IEEE International Conference on Signal Processing and Communications, 2007, pp. 1003-1006.

[3] X. Song, L.Herranz, and S. Jiang, "Depth CNNs for RGB-D scene recognition: learning from scratch better than transferring from RGB-CNNs," *arXiv*:1801.06797 [*cs*, *stat*], Jan. 2018.

[4] Z. Li and N. Snavely, "MegaDepth: Learning Single-View Depth Prediction from Internet Photos," *arXiv*:1804.00607 [*cs*], Apr. 2018

[5] I.-H.Jhuo, S. Gao, L. Zhuang, D. T. Lee, and Y. Ma, "Unsupervised Feature Learning for RGB-D Image Classification," in *Computer Vision -- ACCV* 2014, vol. 9003, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham: Springer International Publishing, 2015, pp. 276-289.

[6] L. Chen, W. Li, and D. Xu, "Recognizing RGB Images by Learning from RGB-D Data," in2014 *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 1418-1425.

[7] S. Wang and Y. Tian, "Detecting stairs and pedestrian crosswalks for the blind by RGBD camera," in2012 *IEEE International Conference on Bioinformatics and Biomedicine Workshops*, 2012, pp. 732-739.

[8] R.Munoz, Xuejian Rong, and Y. Tian, "Depth-aware indoor staircase detection and recognition for the visually impaired," in 2016 *IEEE International Conference on Multimedia &Expo Workshops* (*ICMEW*), Seattle, WA, USA, 2016, pp. 1-6.

[9] Y. He, "Estimated Depth Map Helps Image Classification,"*arXiv*:1709.07077 [*cs*], Sep. 2017.

[10] A. Torralba and A. A.Efros, "Unbiased look at dataset bias," in *CVPR* 2011, 2011, pp. 1521-1528.

[11] Y. Ganin *et al.*, "Domain-Adversarial Training of Neural Networks," in *Domain Adaptation in Computer Vision Applications*, G. Csurka, Ed. Cham: Springer International Publishing, 2017, pp. 189-209.

[12] S. Ben-David, J. Blitzer, K. Crammer, and F. Pereira, "Analysis of Representations for Domain Adaptation," in Advances in Neural Information Processing Systems 19, B.Schölkopf, J. C. Platt, and T. Hoff-

man, Eds. MIT Press, 2007, pp. 137-144.

[13] E. Tzeng, J. Hoffman, N. Zhang, K.Saenko, and T. Darrell. "Simultaneous deep transfer across domains and tasks", In ICCV, 2015.

[14] A.Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097-1105.

[15] Z. Jiang, Y. Wang, L. S. Davis, W. Andrews, and V.Rozgic. "Learning discriminative features via label consistent neural network", In WACV, 2017.

[16] M. D.Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," in Computer Vision-ECCV 2014, vol. 8689, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818-833.

[17] T. He, Z. Zhang, H. Zhang, Z. Zhang, J.Xie, and M. Li, "Bag of Tricks for Image Classification with Convolutional Neural Networks," arXiv:1812.01187 [cs], Dec. 2018.

Authors' Biographies

Jing WANG received the B.A Sc degree from The University of British Columbia (UBC) in 2018. Now he is an M.A Sc student at the University of British Columbia. His research interests include computer vision and intelligent robotics.

E-mail: j.wang94@alumni.ubc.ca

Kuangen ZHANG received the B.E. degree from Tsinghua University in 2016. Now he is a joint Ph.D. student of the University of British Columbia (UBC) and Southern University of Science and Technology (SUSTech). His research interests include robotic vision, sensor fusion, and wearable robots.

E-mail: kuangen.zhang@alumni.ubc.ca