Proceedings of 15th Int. Conf. on Intelligent Autonomous Systems,
Baden-Baden, Germany, June 2018

Convolutional Channel Features-based Person Identification for Person Following Robots

Kenji Koide and Jun Miura

Toyohashi University of Technology, Toyohashi, Aichi, Japan,
koide@aisl.cs.tut.ac.jp, jun.miura@tut.jp

Abstract. This paper describes a novel person identification framework for mobile robots. In this framework, we combine Convolutional Channel Features (CCF) and online boosting to construct a classifier of a target person to be followed. It allows us to take advantage of deep neural network-based feature representation and adapt the person classifier to the specific target person depending on circumstances. Through evaluations, we validated that the proposed method outperforms existing person identification methods for mobile robots. We applied the proposed method to a real person following robot, and it has been shown that CCF-based person identification realizes robust person following.

Keywords: person tracking, person identification, mobile robot

1 Introduction

Person identification is one of the fundamental functions for person following robots. To keep following a person, they have to reliably localize the target person real-time. In cases where the target person is occluded by another person, robots would lose track of him/her, and they have to find the person among surrounding persons with a person model learned before the lost (i.e., re-identification) to resume following.

In this paper, we propose a Convolutional Channel Features (CCF) [21] based person identification framework for person following robots. Fig. 1 shows our person following robot and an overview of the proposed system. The robot is equipped with Laser Range Finders (LRFs) and a camera. We first detect and track people using the LRFs, and then find people regions on images based on the people positions provided by the LRFs. In order to reliably identify a target person among surrounding persons, we construct a target person model by learning his/her appearance. We employ CCF, a set of convolution filters trained by a neural network, to extract appearance features of persons. It produces robust and discriminative features for person identification. Then, the robot learns the extracted features of the target person using online boosting [8]. With this approach, the robot can adapt the person classifier to the specific target

Video available at: https://www.youtube.com/watch?v=semX5Li0yxQ

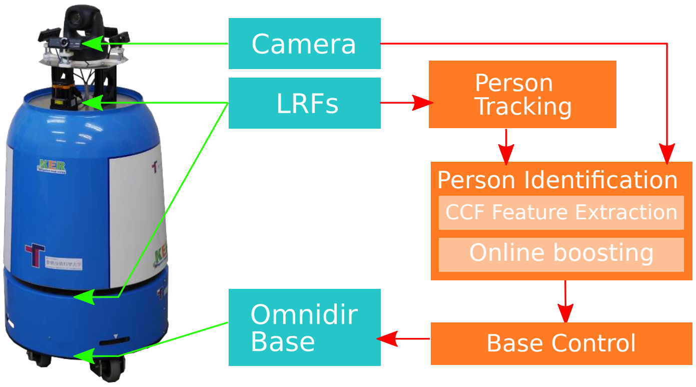


Fig. 1: Person following robot equipped with a camera and LRFs.

person. It is suitable for tasks where a specific person is important, like person following.

The contributions of this paper are two-fold. First, we introduce CCF, deep neural network-based representation, for person following robots. It significantly improves the identification accuracy, while keeping the processing cost low (it can be run real-time without GPU). To our knowledge, this is the first work to introduce it to online person re-identification. Secondly, we provide pieces of the framework, such as CCF related routines and parameters, as open source [1]. It can be re-used for person identification as well as other tasks, such as person detection and tracking.

The rest of the paper is organized as follows. Sec. 2 explains related works. Sec. 3 describes our LRF-based person tracking and visual person region detection methods. Sec. 4 describes the proposed CCF-based person identification method and its evaluation. Sec. 5 shows a person following experiment conducted to show that the proposed method can be applied to real person following robots. Sec. 6 concludes the paper.

2 Related work

Person tracking is an essential function for person following robots. A lot of works proposed person tracking systems for mobile robots. In particular, LRFs [3, 10] and depth cameras [16, 19] have been widely used for person following robots. They provide a person's position accurately as long as he/she is visible from the sensors and promise reliable person following capabilities. However, once the systems lose track of a person to be followed due to occlusion, they cannot find the person even he/she re-appear in the sensor view, and robots are not able to continue to follow the person. In such cases, robots have to re-identify the person based on a target person model learned before the occlusion.

[1] `https://github.com/koide3/ccf_feature_extraction`

Several features, such as gait [14], height [4], and skeletal information [15], have been proposed for person re-identification. In cases of mobile robots, the most popular way is to use appearance features, such as color and texture of cloths, and learn them online [2, 6, 13] to construct a target person model. Appearance is one of the most disciminative features, and online learning methods adapt the person model to a specific target person. For instance, when there are persons wearing similar sheets and dissimilar trousers, online learning methods can focus on the discriminative part, trousers in this case, to re-identify the target person robustly. However, most of existing methods for mobile robots use naive hand-crafted appearance features, such as Haar-like features [13], Local Binary Patterns (LBP) [6], edge features [2] on color and depth images. They are not dedicated features for person re-identification, and they may not be discriminative when persons are wearing similar cloths.

Recently, deep neural networks have been applied to various vision applications. Person re-identification for surveillance is one of such applications, and Convolutional Neural Network (CNN) based methods outperform traditional systems [1, 20] in terms of identification accuracy. However, a few works [5] applied such CNN-based methods to mobile robots due to the limitation of computation resource on mobile robots. We usually cannot use computers with GPUs for mobile robots, and it is hard to directly apply such CNN-based methods to person following robots. Moreover, in person following tasks, it is important to adapt the person model to the target person online. Although there are methods to update neural networks online [18], those methods are very costly.

Yang et al. proposed Convolutional Channel Features [21]. They take the first a few convolution layers from a trained deep CNN, and use the set of convolution layers as a feature extractor (called CCF). By training light-weight models, such as SVM and boosting, with the deep feature representations, they adapt the framework to several tasks without expensive tuning of the network. Following their work, we introduce CCF to person identification for mobile robots to take advantage of deep representation while keeping the processing cost low.

3 Person tracking

3.1 LRF-based person tracking

We first detect and track persons using LRFs placed at leg and torso heights. We detect leg/torso candidate clusters by finding local minimas in range data, and then we validate if they are real torsos/legs using classifiers based on cluster shapes. We use Arras's method (14 shape features and boosting) [3], and Zainudin's method (4 shape features and SVM) [22] to validate torsos and legs, respectively. We assume that the torso and at least one of the legs of a person must be detected, and aggregate the detected torsos and legs by considering torsos with no legs under it are false positives. We track the detected torsos using Kalman filter with a constant velocity model and global nearest neighbor data association [17].

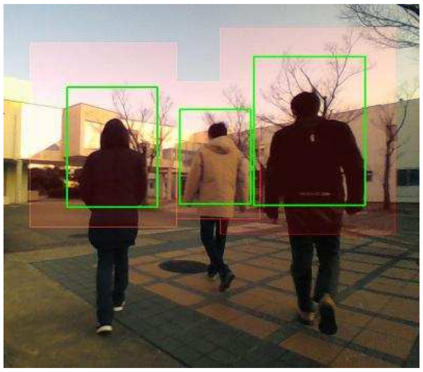


Fig. 2: ROIs calculated from person positions provided by the LRFs (Red transparent regions), and detected upper body regions (Green rectangles).

3.2 Visual person region detection

In order to find person regions on an image for appearance feature extraction, we project a cylinder at each person position into the image, and calculate a rectangle which surrounds the projected cylinder as an ROI. Then, we use HOG cascaded classifier [7] to accurately localize the upper body region of the person. Fig. 2 shows an example of calculated ROIs and detected upper body regions. We double the height of each region so that it covers the whole body of the person, and then extract appearance features from the detected regions to train a target person classifier.

4 Person identification

4.1 Convolutional channel features

To take advantage of deep CNN-based feature representation, we employ Convolutional Channel Features (CCF) [21] instead of traditional appearance features which have been used for mobile robots, such as color histograms [9], haar-like [3], and edge features [2]. CCF consists of a few convolutional layers taken from a trained deep CNN. It takes an input image and yields a set of response maps (i.e. feature maps) which is optimized for a specific task, such as person detection and classification.

In this work, we train Ahmed's network for person re-identification [1] as the base of CCF, and use the first two convolution layers of the network to extract appearance features for online person identification (see Fig.3). Ahmed's network takes a pair of person images and then applies convolution filters to extract feature maps of each input image. The extracted feature maps are compared together by taking the difference between each pixel of a feature map and the neighbor pixels of the corresponding pixel of the other map. Then, it applies convolution filters again to the differences map, and through a linear layer,

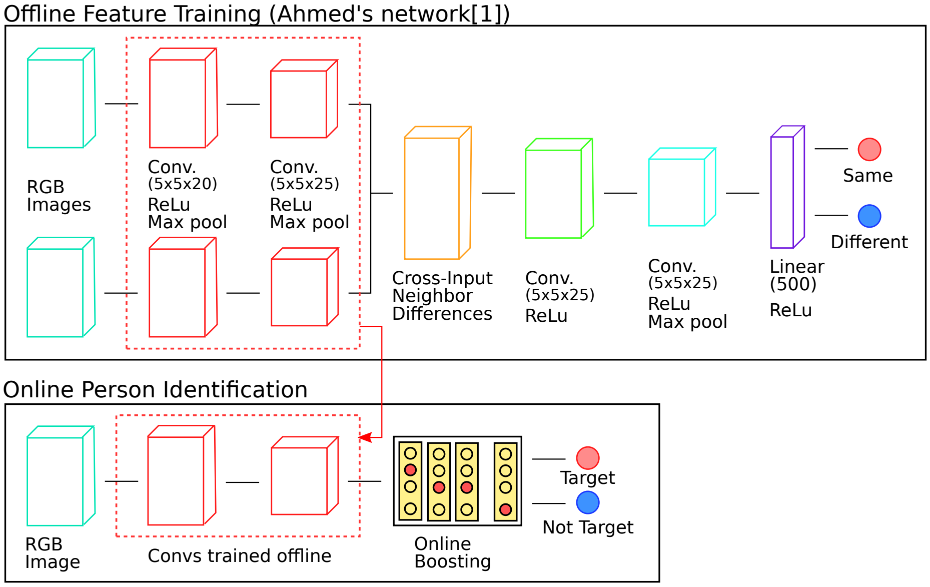

Offline Feature Training (Ahmed's network[1])
RGB Images
Conv. (5x5x20) ReLu Max pool
Conv. (5x5x25) ReLu Max pool
Cross-Input Neighbor Differences
Conv. (5x5x25) ReLu
Conv. (5x5x25) ReLu Max pool
Linear (500) ReLu
Same
Different
Online Person Identification
RGB Image
Convs trained offline
Online Boosting
Target
Not Target

Fig. 3: Convolutional Channel Features-based person identification framework. We take the first two layers of a network for person re-identification and use them to extract features for online person identification.

the network judges whether the input images are the same person or not. The numbers of filters of the first and the second convolution layers in the network are 20 and 25, and thus, they yield 25 feature maps. Since it may be costly for mobile systems to directly use the network, we also trained a tiny version of the network, where the numbers of convolution filters of both the first and the second layers are 10. We trained both the networks with a dataset consisting of CUHK01 [11] and CUHK03 [12]. The total number of identities in the dataset is about 2300, and the number of images is about 17000. We used nine tenths of the dataset for training and the rest for testing and confirmed that both the networks show over 98% of identification accuracies on the test set. In the rest of this paper, the CCFs taken from the original and the tiny version networks are denoted as CCF25 and CCF10, respectively.

Fig. 4 shows example feature maps extracted by CCF10. We can see that each layer shows strong responses for different color properties. For instance, layer 2 shows higher values on darker or blue regions, while layer 8 strongly responds orange regions. We can obtain diverse feature representation using CCF, without hand-crafting, and such diverse features would contribute to identification performances.

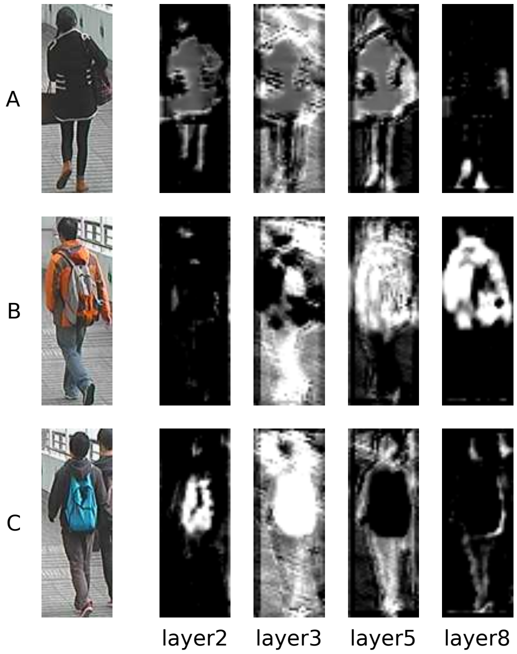
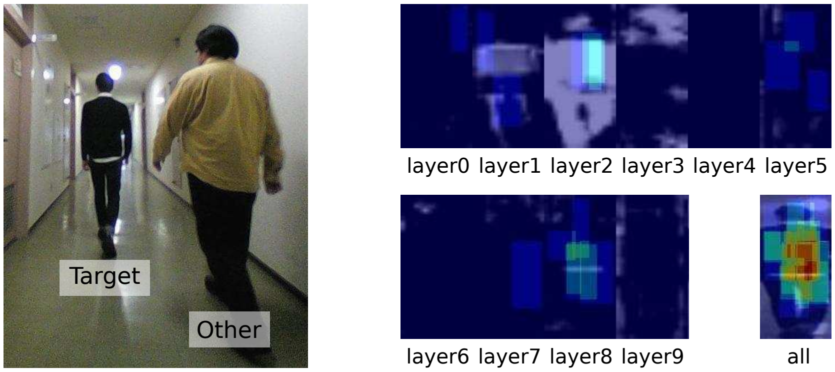


Fig. 4: Feature maps extracted by CCF10. Each layer shows strong responses for different color properties.


Target
Other
layer0 layer1 layer2 layer3 layer4 layer5
layer6 layer7 layer8 layer9
all

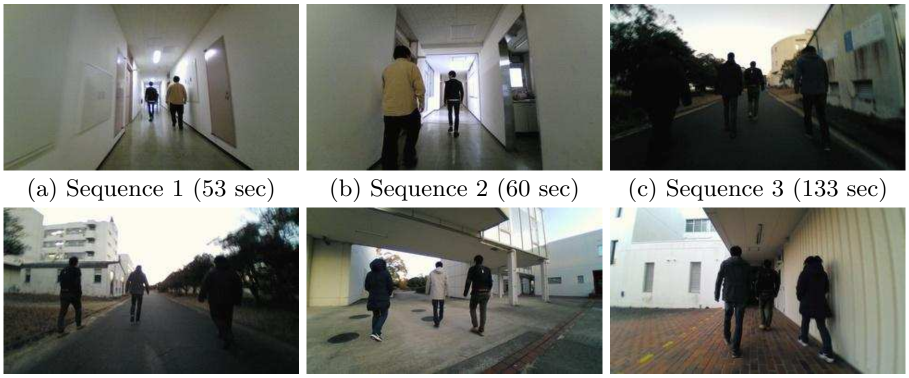
(a) A snapshot of the environment.

(b) Features selected by online boosting. Heatmaps of selected regions are overlayed on feature maps.

Fig. 5: An example of features selected by online boosting. The discriminative regions, the upper body regions in this case, are automatically selected.

4.2 Online boosting-based person classifier

With the offline trained CCF, we extract feature maps from person images, and then train a target person classifier online. Following Luber's work [13], we



(a) Sequence 1 (53 sec) (b) Sequence 2 (60 sec) (c) Sequence 3 (133 sec)

(d) Sequence 4 (175 sec) (e) Sequence 5 (122 sec) (f) Sequence 6 (58 sec)

Fig. 6: Snapshots of the dataset for evaluation of person identification in person following tasks. The dataset consists of RGB images and LRF data recorded from a mobile robot. The robot was manually controlled and following a person in indoor and outdoor environments.

employ online boosting [8] to construct the classifier. Online boosting constructs an ensemble of weak classifiers and uses it as a strong classifier. In this work, each weak classifier takes the sum of pixel values in a random rectangle region on a feature map and classifies images into the target and other persons using a naive Bayes classifier. Since online boosting selects weak classifiers with better accuracies, regions, which are effective to identify the target, are automatically chosen for identification. In this work, we use online boosting with 10 weak classifier selectors, and each selector contains 15 weak classifiers. Thus, the total number of weak classifiers is 150, and 10 of them are selected to construct an ensemble. Fig. 5 shows an example of features selected by online boosting. We can see that online boosting automatically selects the discriminative regions, the upper body regions, in this case, to construct a classifier ensemble.

4.3 Evaluation

To evaluate the proposed CCF-based person identification method, we created a dataset consisting RGB image and LRF data sequences taken from a person following robot (shown in Fig. 1). Fig. 6 shows snapshots of the dataset. We controlled the robot manually and made it follow a target person in indoor and outdoor environments. We collected six sequences, and two of them are recorded in indoor, and the rest are recorded in outdoor environments. In each sequence, a target person to be followed stands in front of the robot for the first seconds so the robot can know and learn the appearance of the person, and then he/she starts walking. During the recording, the target person was often occluded by other persons so he/she become invisible from the robot, and the robot loses track of him/her.

Table 1: Person identification evaluation result. Bold indicates best results.

		Duration [sec]		
		Haar Lab	CCF10	CCF25
Seq. 1	CT	38.78 (73.23%)	**40.84 (77.11%)**	37.96 (71.69%)
	CL	6.62 (12.49%)	6.78 (12.80%)	**7.37 (13.92%)**
	WT	3.91 (7.38%)	3.75 (7.08%)	**3.16 (5.96%)**
	WL	3.65 (6.90%)	**1.59 (3.01%)**	4.47 (8.44%)
Seq. 2	CT	43.76 (73.78%)	43.86 (73.95%)	**43.87 (73.97%)**
	CL	**11.28 (19.02%)**	10.76 (18.14%)	10.90 (18.37%)
	WT	**2.52** (4.24%)	3.04 (5.12%)	2.90 (4.89%)
	WL	1.76 (2.96%)	1.65 (2.79%)	**1.64 (2.77%)**
Seq. 3	CT	48.08 (36.11%)	**106.31 (79.84%)**	88.60 (66.55%)
	CL	7.67 (5.76%)	**20.18 (15.16%)**	19.67 (14.77%)
	WT	46.45 (34.89%)	**3.94 (2.96%)**	6.47 (4.86%)
	WL	30.94 (23.24%)	**2.71 (2.04%)**	18.40 (13.82%)
Seq. 4	CT	37.89 (21.56%)	**141.19 (80.33%)**	85.60 (48.70%)
	CL	**24.83 (14.13%)**	23.18 (13.19%)	21.57 (12.27%)
	WT	12.08 (6.88%)	**5.83 (3.32%)**	6.30 (3.58%)
	WL	100.95 (57.44%)	**5.56 (3.16%)**	62.29 (35.44%)
Seq. 5	CT	98.33 (80.38%)	98.75 (80.73%)	**98.89 (80.84%)**
	CL	16.66 (13.62%)	**18.39 (15.03%)**	18.36 (15.00%)
	WT	5.12 (4.19%)	**3.32 (2.71%)**	3.38 (2.76%)
	WL	2.22 (1.81%)	1.88 (1.53%)	**1.70 (1.39%)**
Seq. 6	CT	33.10 (59.67%)	41.90 (75.55%)	**43.67 (78.74%)**
	CL	2.68 (4.84%)	**9.01 (16.24%)**	**9.01 (16.24%)**
	WT	16.80 (30.28%)	**0.06 (0.11%)**	**0.06 (0.11%)**
	WL	2.88 (5.20%)	4.49 (8.10%)	**2.73 (4.91%)**
Total	CT	299.94 (50.08%)	**472.86 (78.94%)**	398.60 (66.55%)
	CL	69.75 (11.64%)	**88.29 (14.74%)**	86.87 (14.50%)
	WT	86.89 (14.51%)	**19.94 (3.33%)**	22.26 (3.72%)
	WL	142.40 (23.77%)	**17.89 (2.99%)**	91.24 (15.23%)

Table 2: Processing time for each person image

	method	time [msec]
feature extraction	Haar & Lab	1.2
	CCF10	4.2
	CCF25	6.0
classifier update	all	0.1

We evaluated the proposed method with CCF25 and CCF10 on the dataset. We also evaluated online boosting with Haar-like features on intensity images and *Lab* color histograms. This is almost identical to [13] except that we didn't use Haar-like features on depth images.

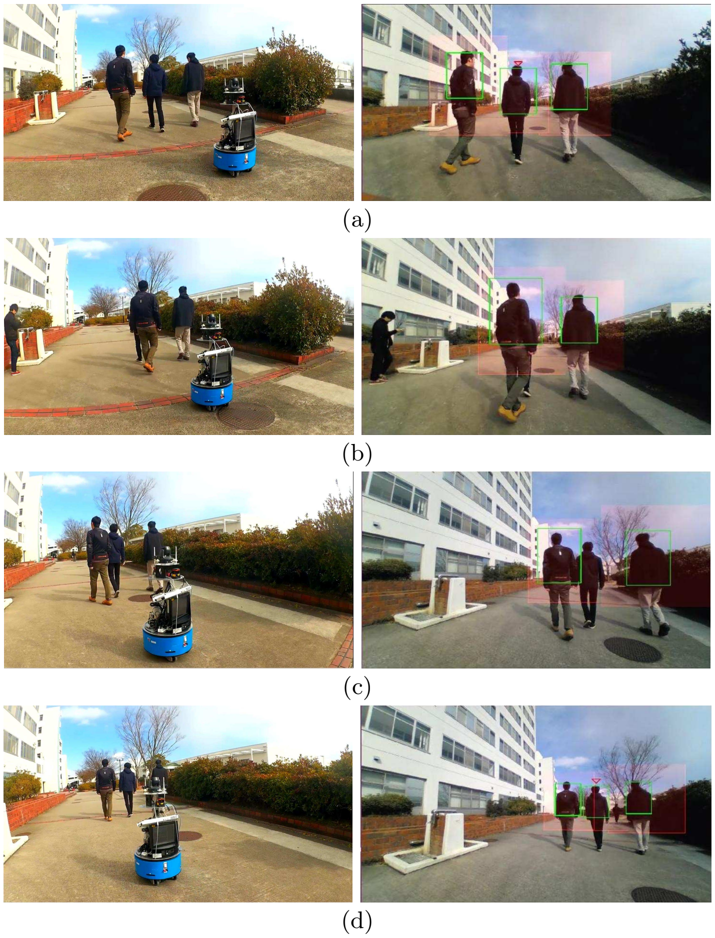
Table 1 shows a summary of identification results. We categorized identification results in four states. CT (Correctly Tracked) means that the target was visible from the robot and correctly identified. CL (Correctly Lost) means that the target was invisible from the robot due to occlusion, and the system judged that he/she is not in the view correctly. WT (Wrongly Tracked) means the robot identified a wrong person as the target while the target was invisible, and WL (Wrongly Lost) means the robot judged that the target is not visible, although he/she was actually visible from the robot.

CCF-based methods show better identification performance than the traditional appearance feature-based method thanks to their robust deep feature representations. Even in sequences where cloths of the target and others are similar, they correctly identified the target while the traditional one identified wrong persons as the target.

CCF10 and CCF25 show comparable results. However, in a few sequences, CCF25 failed to keep identifying the target person. For instance, it identified a wrong person as the target in sequence 3 and failed to re-identify the target after occlusion in sequence 4. We consider that this is due to the limitation of feature selection of online boosting. Online boosting selects better classifiers among a limited number of weak classifiers. When the feature space is vast, the set of weak classifiers cannot cover enough feature space, and thus, online boosting would fail to select good features. The performance of CCF25 could be improved by increasing the number of weak classifiers. However, it increases the processing cost, and it may lead to over-fitting. Although the feature space of CCF10 is smaller than CCF25, “average effectiveness” of CCF10 features might be better than CCF25 since it was optimized to identify persons with fewer filters. As a result, CCF10 showed a better result than CCF25 in this case.

Note that, we also tested the original Ahmed’s network on this dataset, however, the results were very poor. In each sequence, we compared every person image with the target person images of the first ten seconds using the network, and classified the image into the target and others by majority-voting. However, it worked well on only easy situations (Sequence 1 and 2), and in the rest of sequences, it classified all similar persons as the target (Sequence 3, 4, and 6) or classified the target as other persons (Sequence 5). The results suggest that, even with deep feature representations, we cannot obtain a good identification result without online learning approaches. In addition to that, it took about 1 sec for each frame and was far from real-time performance.

Table 2 shows average processing time of feature extraction and person classifier update on a computer with Core i7-6700K (without GPU). While the traditional feature extraction method takes 1.2 msec for each person image, CCF10 and CCF25 take 4.2 msec, and 6.0 msec, respectively. Although CCFs are more costly than the traditional one, they are still able to run real-time. Since the processing time of updating the person classifier depends on only the number of weak classifiers, every method takes the same time for updating (0.1 msec per person image).



(a)

(b)

(c)

(d)

Fig. 7: A person following experiment. The left images are the snapshots of the experiment, and the right images are the detection and the identification results. The red triangles in the right images indicate the person identified as the target.

5 Person following experiment

To show that the proposed method can be applied to real person following tasks, we conducted a person following experiment. We implemented a simple person following strategy; the robot moves toward the target person, and when the robot loses track of the target, it stops and waits until the person re-appears.

Fig. 7 shows snapshots of the experiment. The target person was occluded by another person (b), and the robot lost track of him (c). However, when the target person re-appeared, the robot correctly re-identified him with the boosting model trained before the occlusion and resumed to follow him (d). The duration of the experiment was 220 sec. During the experiment, the robot lost the track

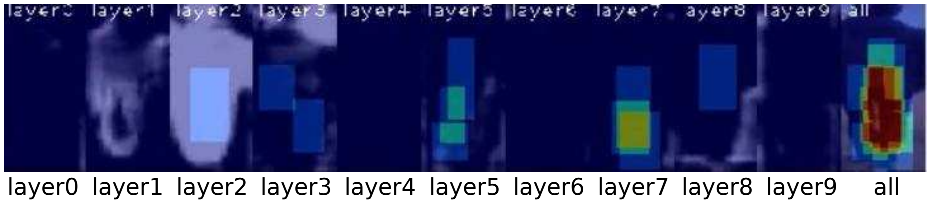


Fig. 8: Features selected by online boosting during the person following experiment.

of the target person seven times due to occlusion. Every time the target person re-appeared in the view of the robot, the robot correctly re-identified the target person and resumed to follow him. Fig. 8 shows the features selected by online boosting during the experiment. Since, in this experiment, persons were wearing jackets with similar colors and trousers with different colors, online boosting selected features around the trousers region.

In this experiment, we used Intel NUC with Core i7-5557U for not only person identification but also other components required to drive the robot. Although the processor is not very powerful one, the system run real-time (about 10Hz).

6 Conclusion

We proposed a novel person identification framework for mobile robots. It is based on Convolutional Channel Features-based appearance features and online boosting to reliably identify the target person to be followed.We validated that the proposed method outperforms a traditional person identification method through evaluations. We also applied the proposed method to a real person following task. It has been shown that the robot with the proposed method is able to robustly follow a target person for a long time.

References

1. Ahmed, E., Jones, M., Marks, T.K.: An improved deep learning architecture for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3908–3916. IEEE (2015)
2. Alvarez-Santos, V., Pardo, X.M., Iglesias, R., Canedo-Rodriguez, A., Regueiro, C.V.: Feature analysis for human recognition and discrimination: Application to a person-following behaviour in a mobile robot. Robotics and autonomous systems 60(8), 1021–1036 (2012)
3. Arras, K.O., Mozos, O.M., Burgard, W.: Using boosted features for the detection of people in 2D range data. In: IEEE International Conference on Robotics and Automation. pp. 3402–3407. IEEE (2007)
4. Berdugo, G., Soceanu, O., Moshe, Y., Rudoy, D., Dvir, I.: Object reidentification in real world scenarios across multiple non-overlapping cameras. In: European Signal Processing Conference. pp. 1806–1810 (2010)

5. Chen, B.X., Sahdev, R., Tsotsos, J.K.: Integrating stereo vision with a CNN tracker for a person-following robot. In: Lecture Notes in Computer Science, pp. 300–313. Springer International Publishing (2017)
6. Chen, B.X., Sahdev, R., Tsotsos, J.K.: Person following robot using selected online ada-boosting with stereo camera. In: Conference on Computer and Robot Vision. pp. 48–55 (2017)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 886–893. IEEE (2005)
8. Grabner, H., Bischof, H.: On-line boosting and vision. In: IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 260–267. IEEE (2006)
9. Koide, K., Miura, J.: Identification of a specific person using color, height, and gait features for a person following robot. Robotics and Autonomous Systems 84, 76–87 (2016)
10. Leigh, A., Pineau, J., Olmedo, N., Zhang, H.: Person tracking and following with 2d laser scanners. In: IEEE International Conference on Robotics and Automation. pp. 726–733 (2015)
11. Li, W., Zhao, R., Wang, X.: Human reidentification with transferred metric learning. In: Asian Conference on Computer Vision (2012)
12. Li, W., Zhao, R., Xiao, T., Wang, X.: Deepreid: Deep filter pairing neural network for person re-identification. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
13. Luber, M., Spinello, L., Arras, K.O.: People tracking in RGB-d data with online boosted target models. In: IEEE/RSJ International Conference on Intelligent Robots and Systems. pp. 3844–3849. IEEE (2011)
14. Makihara, Y., Mannami, H., Yagi, Y.: Gait analysis of gender and age using a large-scale multi-view gait database. In: Asian Conference on Computer Vision, pp. 440–451. Springer (2011)
15. Munaro, M., Ghidoni, S., Dizmen, D.T., Menegatti, E.: A feature-based approach to people re-identification using skeleton keypoints. In: IEEE International Conference on Robotics and Automation. pp. 5644–5651. IEEE (2014)
16. Munaro, M., Menegatti, E.: Fast RGB-d people tracking for service robots. Autonomous Robots 37(3), 227–242 (2014)
17. Radosavljevic, Z.: A study of a target tracking method using global nearest neighbor algorithm. Vojnotehnicki glasnik (2), 160–167 (2006)
18. Sahoo, D., Pham, Q., Lu, J., Hoi, S.C.H.: Online deep learning: Learning deep neural networks on the fly. CoRR abs/1711.03705 (2017), http://arxiv.org/abs/1711.03705
19. Satake, J., Chiba, M., Miura, J.: A SIFT-based person identification using a distance-dependent appearance model for a person following robot. In: IEEE International Conference on Robotics and Biomimetics. pp. 962–967. IEEE (2012)
20. Schumann, A., Stiefelhagen, R.: Person re-identification by deep learning attribute-complementary information. In: IEEE Conference on Computer Vision and Pattern Recognition Workshops. IEEE (2017)
21. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Convolutional channel features. In: IEEE International Conference on Computer Vision. IEEE (2015)
22. Zainudin, Z., Kodagoda, S., Dissanayake, G.: Torso detection and tracking using a 2d laser range finder. In: Australasian Conference on Robotics and Automation. ARAA (2010)