IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 28, 2021, accepted July 29, 2021. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.3104973

Aggregating Everyday Outfits by Incremental Clustering With Interactive User Adaptation

YASUTOMO KAWANISHI[1,2,3], (Member, IEEE), HIROSHI MURASE[2,3], (Life Fellow, IEEE), SATOSHI KOMORITA[3], AND SEI NAITO[3], (Member, IEEE)

[1] Multimodal Data Recognition Research Team, Guardian Robot Project, R-IH, RIKEN, Kyoto 619-0288, Japan
[2] Graduate School of Informatics, Nagoya University, Nagoya, Aichi 464-8601, Japan
[3] KDDI Research, Saitama 356-8502, Japan

Corresponding author: Yasutomo Kawanishi (yasutomo.kawanishi@riken.jp)

This work was supported in part by the MEXT (Ministry of Education, Culture, Sports, Science and Technology, JAPAN) through Grant-in-Aid for Scientific Research under Grant JP17H00745.

ABSTRACT Knowledge of the outfits that a person wears daily and how frequently the person wears them will help the person select clothing every morning. However, it is very-time consuming to manually record what the person wears every day. This paper proposes a system that automatically aggregates and visualizes the outfits of a user by using a monitoring camera at home. To aggregate the everyday outfits of a user, we employ incremental clustering. For accurate clustering, an appropriate feature space is required. However, there is a gap between the clothing feature space of people and a specific user. To fill the gap, we propose a Siamese-network based interactive user adaptation method using user feedback. The user adaptation incrementally updates the similarity metric of the clothing feature space. We confirmed that the proposed system achieves highly accurate clustering performance with a smaller amount of user feedback through evaluation.

INDEX TERMS Clothing frequency, fashion application, incremental clustering, metric learning, user adaptation, user feedback.

I. INTRODUCTION

Choosing the appropriate clothes to wear every morning is an arduous task for many of us. Various factors such as avoiding a repetition of the same clothes on consecutive days and wearing one's favorite clothes frequently are taken into account while choosing one's outfit. In addition, suitable combinations need to be considered when buying new clothes. Fashion applications aid users in making such decisions.

Fashion application development is an emerging research field, and fashion recommendation systems have become a topic of significant interest [1]–[5]. However, a recommendation system has to consider many variables such as the types of clothes the user owns, how frequently these clothes are worn, and personal preferences to satisfy the above requirements.

If everyday outfits are recorded manually by a user, the frequency of choosing to wear the same clothes can be ascertained. However, it is a tough task to maintain this record every morning, especially with hectic schedules. To the best of our knowledge, there is no system to record this data automatically.

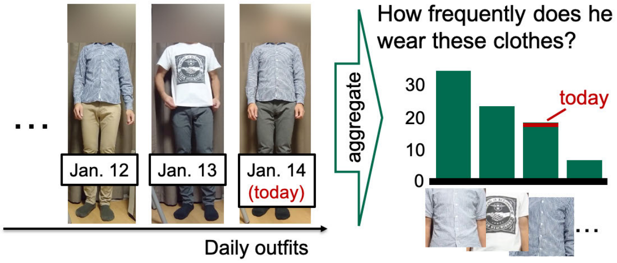


FIGURE 1. Example of an output of the proposed system.

In this study, we propose a system that automatically "aggregates" the everyday outfits of users and displays how frequently a user wears each outfit (Fig. 1) without manual recording. Here, the term aggregate stands for collecting pieces of clothing and count each of them. The system shows users the aggregation results whenever it captures a piece of clothing, and the users can correct the results manually in case of misrecognition. We call this correction user feedback. Recently, smart speakers, such as Google Home [6], have seen increased usage, and some of the systems such as Echo

The associate editor coordinating the review of this manuscript and approving it for publication was Hiu Yung Wong.

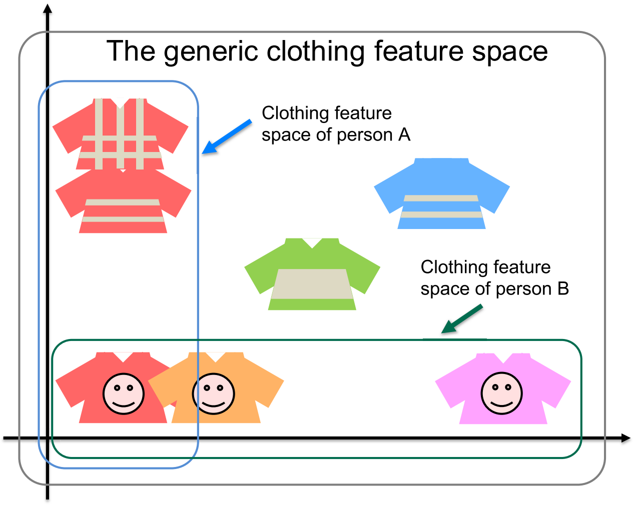

The generic clothing feature space
Clothing feature space of person A
Clothing feature space of person B

FIGURE 2. Example of the gap between the generic clothing feature space and user-specific clothing feature spaces.

Look [7] have in-built cameras. It is expected that cameras installed in smart speakers will become more popular in the coming days. In the proposed system, we assume that users are observed every morning by such cameras mounted in their rooms, and these observations are used as input images for the proposed system. Surveillance cameras that are installed in entrances or rooms can also be used.

By applying the incremental clustering algorithm [8] to detected clothing images from a camera, the system can aggregate clothing and count them every day. For reducing the frequency of user feedback, the clustering should be highly accurate. For accurate clustering, appropriate image features that have discriminative power are expected to be extracted from clothing images. A discriminative feature space can be trained using a large number of diverse clothing images. We call the feature space *the generic clothing feature space*. However, clothing preferences vary from person to person; the clothing variation of a user is much smaller than the clothing variations in the world. Therefore, there is a gap between the clothing feature space of each specific user and the generic clothing feature space, as shown in Fig. 2. It would cause the generic clothing feature space not to be appropriate for each user.

To fill the gap, we introduce an interactive user adaptation method based on user feedback. By fine-tuning a feature extractor using user feedback, the generic clothing feature space will be transferred to *the user-specific clothing feature space*. Better clustering can be expected in the user-specific clothing feature space. In the proposed system, we assume that this procedure is run on a cloud server whenever the user feedback is provided; thus, the clothing feature space is updated on demand.

The contributions of this study can be summarized as follows:

- **Everyday outfits aggregation system**: we propose an automatic aggregation system of clothing images every day based on incremental clustering with user feedback.
- **Interactive User Adaptation**: we propose an interactive user adaptation method that adapts the generic clothing feature space to a user-specific clothing feature space based on user feedback.
- **Evaluation using a real clothing dataset**: we collected a dataset of about 7,000 images by taking photos of twelve users every morning for a month.

To simplify the discussion, in this paper, we assume that the system only aggregates the upper part of the human body. However, the system can be easily extended to a full-body aggregation system.

The rest of this paper is organized as follows: In Section II, recent work on fashion analysis and recommendation is summarized. In Section III, the details of the proposed clothing aggregation system are discussed. In Section IV, the proposed interactive user adaptation method is explained in detail. Experimental results are presented in Section V. Finally, we conclude the paper in Section VI.

II. RELATED WORK

Fashion analysis and recommendation are extensively studied topics in the field of information retrieval, computer vision, and image processing [9]–[13]. Recently, the usage of clothing recommendation systems [2], [14]–[16] that match users' preferences has become a trend on e-commerce websites. Yu *et al.* [15] have proposed a recommendation system that focuses on aesthetics, which is highly relevant to user preference. Abe *et al.* [17] collected large datasets of images from YFCC100M [18] and have analyzed clothing trends with regard to geolocation, especially for metropolitan areas. The method proposed by He and McAuley [19] also focuses on the trend of fashion.

Many studies have focused on selecting clothing at home, and several user interfaces to support selecting clothing at home have been proposed [20], [21]. When a user is selecting a piece of clothing, the systems show several recommending pieces of clothing to the user. McAuley *et al.* [22] proposed a recommendation system using a query image. Tangseng *et al.* [23] proposed a recommendation system for clothing in a closet. These existing clothing recommendation systems require registering all the clothing that the user owns to the closet database and also requires recording everyday outfits manually. This is a time-consuming process.

Recently, the fashion generation is also been a hot topic in this field [24]–[29]. There are various methods to generate fashion images by using GANs that have been proposed. For training a GANs model and generate user-specific fashion images, it is also required to collect a large-scale everyday outfits of the user beforehand.

III. THE PROPOSED CLOTHING AGGREGATION SYSTEM

A. OVERVIEW

The goal of the proposed system is to automatically aggregate clothing items usually worn by a user and to determine how frequently the user wears them. By incrementally clustering

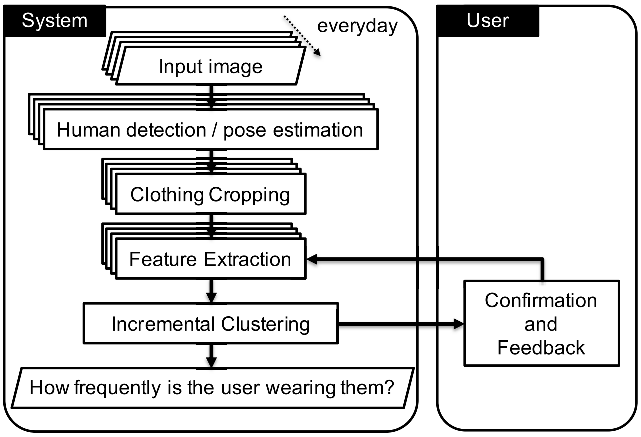

System
everyday
Input image
Human detection / pose estimation
Clothing Cropping
Feature Extraction
Incremental Clustering
How frequently is the user wearing them?
User
Confirmation and Feedback

FIGURE 3. Process flow of the proposed system: feature extraction and incremental clustering are applied to cropped images. The user provides corrections as user feedback if the clustering results are incorrect.

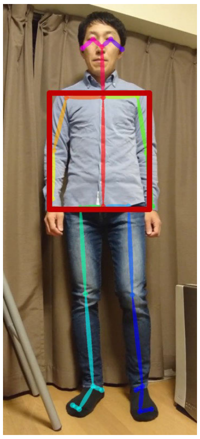


FIGURE 4. Example result of human pose estimation: the red rectangle indicates the extracted part of the upper body of the person.

clothing photos taken every morning, the system realizes the aggregation and calculates the frequency. The process flow of the proposed method is shown in Fig. 3. We assume that each photo is taken at the user's room entrance every morning when the user goes out. Therefore, we can assume that the photos are taken in the same environment, and the users in the photos are in similar postures.

B. HUMAN DETECTION AND POSE ESTIMATION

Human detection and pose estimation are applied to extract the regions of clothing from photos. As a result, the locations of the joints of the human body are obtained. An example result of the pose estimation is shown in Fig. 4.

Although there is a possibility that multiple people can be detected in a photo, we assume that only one person is detected in all photos. It is because we assume the camera is installed at the user's room entrance.

C. CROPPING OF THE CLOTHING REGION FROM AN IMAGE

The proposed system crops clothing regions based on the locations of body joints obtained by the process described in Section III-B.

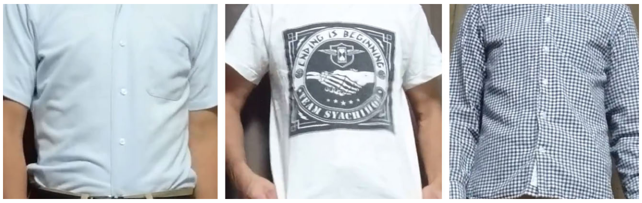


FIGURE 5. Examples of cropped images.

For example, given the human pose, an image of clothing for the upper body is cropped based on the minimum bounding box of the body joints of the upper body. We define the upper body by the locations of shoulders, elbows, and hip. For a lower body image, we use the locations of hip, knees, and ankles. Examples of the cropped images are shown in Fig. 5. The following processes are applied independently to these upper and lower body images.

D. FEATURE EXTRACTION

To cluster clothing images accurately, we need an appropriate feature space. Here, we build a discriminative feature space using diverse kinds of clothing, named the generic clothing feature space. We train a feature extractor that maps an input image to a vector in the generic clothing feature space.

We use a convolutional neural network (CNN) as the image feature extractor. Pre-trained ResNet [30] models are often used as the backbones of feature extractors. The pre-trained ResNet model is trained using the ImageNet dataset [31], which consists of various images of a large number of classes. Therefore, it is not suitable for capturing clothing features accurately as is. Therefore, we combine a trainable fully-connected (FC) layer that outputs a p-dimensional vector with the ResNet. We use the p-dimensional vectors as generic clothing features. By using the feature extractor f, given an image I_t, we can extract a generic clothing feature $\mathbf{f}_t \in \mathbb{R}^p$ as

$$\mathbf{f}_t = f(I_t; \widehat{\Theta}), \tag{1}$$

where $\widehat{\Theta}$ is a set of trained parameters of the feature extractor.

The purpose of the feature extraction here is clothing clustering. Thus, we train the model as metric learning to tune the similarity metric of the features.

For metric learning, we employ the contrastive loss [32] with a Siamese network, which is widely used in person re-identification [33]–[35]. The clothing feature space is optimized by using the loss. The loss calculation is shown in Fig. 6. When two given pieces of clothing are the same instance, they should have similar features, and when two given pieces of clothing are different, they should have different features. The contrastive loss using the cosine similarity $s(\cdot, \cdot)$ is defined as

$$L(\mathbf{f}_1, \mathbf{f}_2) = \begin{cases} 1 - s(\mathbf{f}_1, \mathbf{f}_2), & \text{(same)} \\ \max(M - (1 - s(\mathbf{f}_1, \mathbf{f}_2)), 0), & \text{(different)} \end{cases} \tag{2}$$

$$s(\mathbf{f}_1, \mathbf{f}_2) = \frac{\mathbf{f}_1^\top \mathbf{f}_2}{\|\mathbf{f}_1\| \|\mathbf{f}_2\|}, \tag{3}$$

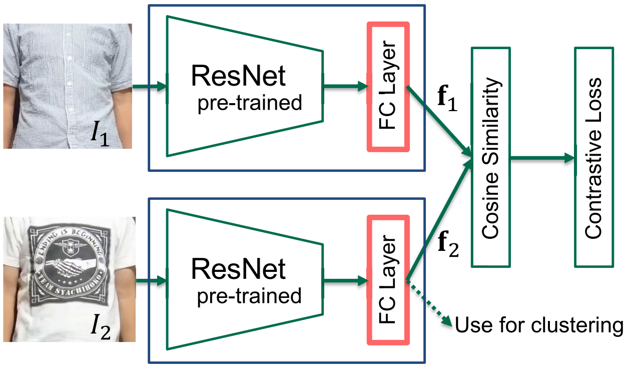


FIGURE 6. Contrastive loss calculation. Given two image features extracted by the CNN independently, the contrastive loss is calculated. Parameters of the last FC layer indicated by the red rectangle are optimized while those of the Resnet is frozen. The output of the last FC layer will be used as a generic clothing feature.

where M denotes the margin parameter. In the training phase, the similarities of pairs of the same clothing should be larger as large as possible, and the similarities of pairs of different clothing should be smaller than the margin M. Here, M is empirically determined.

E. INCREMENTAL CLUSTERING OF CLOTHING IMAGES

The system captures a photo every day, crops a clothing image of the user, and stores the clothing image to the database. The number of clothing items remains unknown, and it will change when the user buys new clothes or discards old ones. Therefore, we use the incremental clustering algorithm [8] in the system.

In incremental clustering, the set of cluster centers C is initialized as an empty set ($C = \emptyset$). When a new clothing image I_t is stored in the database, the similarity $s(\mathbf{f}_t, \mathbf{c}_i)$ between the image feature $\mathbf{f}_t$ of the photo I_t and each cluster center $\mathbf{c}_i \in C$ is calculated. Then, the cluster whose similarity to the input is the maximum is selected as

$$i_t = \arg\max_i s(\mathbf{f}_t, \mathbf{c}_i). \tag{4}$$

When the maximum similarity $s_{i_t} = s(\mathbf{f}_t, \mathbf{c}_{i_t})$ is greater than the margin M, which is as same as the M in the previous section, the input is added to the cluster i_t, and the center of the cluster is updated by using the average of the samples in the cluster (Fig. 7 (i)). Furthermore, when the set of the cluster centers C is empty, or when the maximum similarity s_{i_t} is less than or equal to the margin M, the input is considered as a new piece of clothing and registered as a new cluster (Fig. 7 (ii)).

Finally, the system counts the number of samples for each cluster and outputs the results.

IV. PROPOSED INTERACTIVE USER ADAPTATION

As discussed in Section I, since there is a gap between the generic clothing feature space and appropriate feature space for a specific user, we introduce an interactive user adaptation using user feedback. By utilizing the feedback, the generic clothing feature space will be transferred to the user-specific clothing feature space.

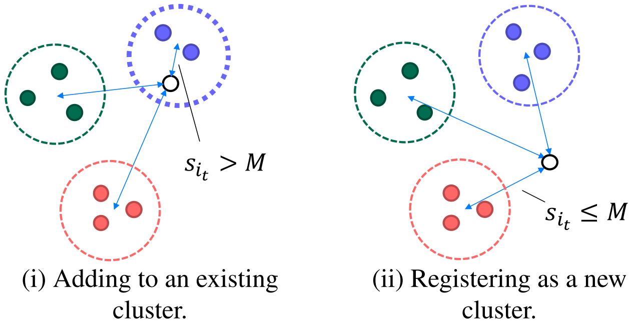

$s_{i_t} > M$
$s_{i_t} \leq M$
(i) Adding to an existing cluster.
(ii) Registering as a new cluster.

FIGURE 7. Examples of the incremental clustering procedure: the small white circle indicates an incoming data, and colored ones indicate existing data. The colored dashed circles indicate existing clusters.

A. OVERVIEW

The generic clothing feature space is defined by metric learning, as described in Section III-D. In the generic clothing feature space, diverse clothing features can be discriminated against. However, the set of clothing of users is small subsets of the diverse clothing. Owing to preference, clothing items of a user are often similar. Therefore, their features would distribute locally in the generic clothing feature space, and it is difficult to distinguish the features of the clothing of the specific users in the generic clothing feature space. Fig. 2 shows an example of the relation between the generic clothing feature space and the clothing feature spaces of several users. In Fig. 2, person A prefers red clothing, and person B prefers clothing face printed on them. Both of these sets are small subsets of the clothing in the world.

To fill the gap between the generic clothing feature space and a user-specific clothing feature space, we utilize user feedback, as shown on the right side of Fig. 3, to update the metric of the feature space.

B. USER FEEDBACK

In the process flow of the clothing aggregation system (Fig. 3), whenever the system receives an input image and performs clustering, users are expected to confirm the clustering results. If a user finds that the label assigned to the input of the day is wrong, he/she is expected to modify the result to the correct label. Since the clustering algorithm is the incremental clustering described in Section III-E, the user only needs to confirm the label of the latest input. If the user modifies the clustering result every day, all of the results will be stored correctly.

For modifying the wrong label, the system shows the image of each cluster center, and then the user selects the correct cluster that the current clothing should belong to. If there is no cluster for the current clothing, the user needs to create a cluster for it.

Since confirming whether the assigned label is correct or not (confirmation) is an easy task, it would be acceptable for users. On the other hand, since selecting the correct label (modification) is a cumbersome task, it should be a rare event. The frequency of label correction depends on the performance of the clustering method. Therefore, user adaptation is performed to improve clustering performance.

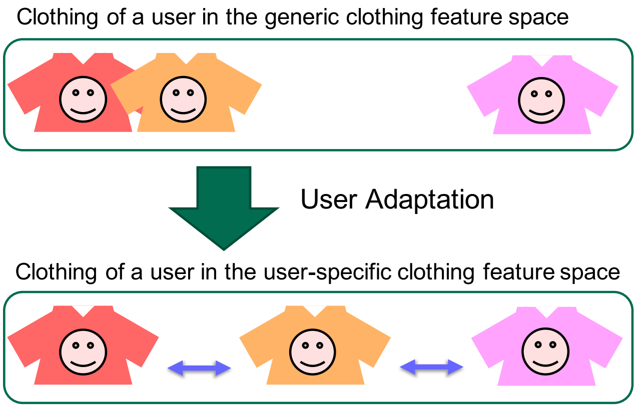

Clothing of a user in the generic clothing feature space
User Adaptation
Clothing of a user in the user-specific clothing feature space

FIGURE 8. Example of the domain adaptation via metric update.

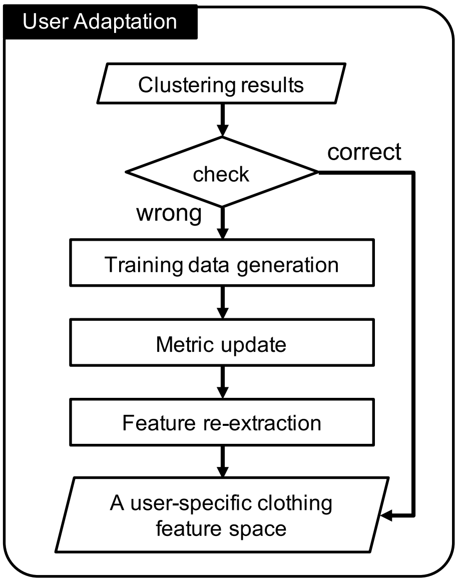


FIGURE 9. The process flow of the proposed user adaptation.

C. FEATURE EXTRACTOR UPDATING AND FEATURE RE-EXTRACTION

We propose a method that transforms the feature space to a user-specific clothing feature space to distinguish the clothing of users clearly (Fig. 8).

For transforming the generic clothing feature space to a user-specific clothing feature space, the feature extractor is modified by using user feedback. Since the label assignment is expected to be corrected by the user feedback, we use the assigned labels by the incremental clustering as the ground truth for updating the feature extractor. The process flow of the proposed user adaptation is shown in Fig. 9.

Since the feature extractor is trained based on the Siamese network explained in Section III, we need a set of pairs of clothing as training data for the user adaptation procedure. K pairs of clothing in the same cluster are randomly sampled, and K pairs of clothing in different clusters are also randomly sampled. In case they are sampled from the same cluster, the features of the two pieces of clothing should be similar. On the other hand, in case they are sampled from different clusters, their clothing features should be different. The feature extractor is updated by using the $2K$ pairs. For updating the model (Fig. 6), backpropagation is performed over the $2K$ images. This makes the feature extractor fit to the set of clothing of the user, and the similarity metric in the feature space is optimized.

Here, we assume that the system stores all of the cropped clothing images on a cloud server. After updating the feature extractor, all features of the clothing of the user are updated by re-extracting the features using the feature extractor. This update procedure is run on the cloud server whenever a user feedback is provided.

V. EVALUATION

A. DATASET

To evaluate the proposed system, we needed a dataset containing images of the everyday outfits of several persons. To the best of our knowledge, there are no such public datasets; hence, we collected a dataset of photos of everyday outfits. The dataset contains photos of twelve subjects. Each subject captured photos of their own outfits every day for about a month, while they selected their own outfits, as usual everyday. The outfits of each subject are selected by themselves everyday. Here, we assume that the subjects do not change their outfits within a day. The numbers of pieces of their outfits vary from 9 to 26. We assume that a camera is installed at the entrance of the room of each subject. Therefore, the orientations of subjects are similar, and the background and the illumination are almost the same within photos of each user. Whenever the subjects captured their outfits, they took about twenty photos in different postures. The total number of photos is 7,456. Samples of the photos are shown in Fig. 10.

B. EVALUATION SETTINGS

Since the clustering results depend on the order of the input sequence, we simulated that a subject selects a piece of clothing every day for D days by using the captured dataset. We randomly sampled D images without duplication for each subject and made a sequence of length D for each subject. We performed the sampling of sequences ten times per subject randomly. Finally, we prepared $120(= 12\times10)$ sequences of clothing.

We performed leave-one-person-out cross-validation. For each subject, we trained the generic clothing feature space using the clothing of the rest of the subjects. The weights of the ResNet were then frozen, and only the weights of the FC layers were updated. Here, ResNet-50 is used as the backbone.

For human body detection and pose estimation, we used OpenPose, proposed by Cao *et al.* [36]. For feature extraction, we set the dimension of feature vectors as $p = 512$. For the user adaptation, we empirically used $M = 0.9$ and applied backpropagation one epoch for each user feedback

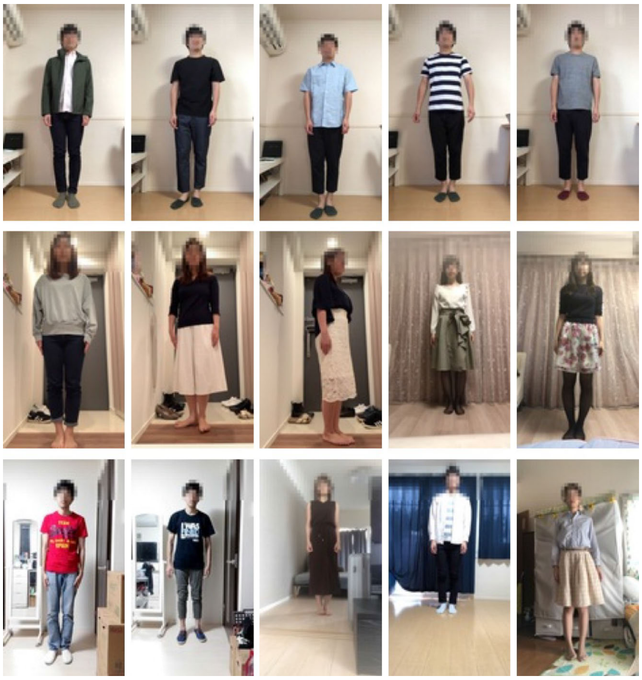


FIGURE 10. Samples of the photos used in the evaluation.

in the evaluation. We used Adam [37] is used with a learning rate 0.001.

As an evaluation metric for clustering, we used the adjusted Rand index (ARI) [38], which is commonly used for evaluating clustering results. The ARI is a metric that compares two label assignments. Here, we assume that the assigned labels by clustering algorithms and the ground truth labels are given for each photo. The ARI score achieves the highest value 1.0, when both of the label assignments are the same, and it can be lower than zero if the label assignments are almost different. Let us assume label sets $X = \{X_1, X_2, \ldots, X_A\}$ and $Y = \{Y_1, Y_2, \ldots, Y_B\}$ are assigned to N elements, and let n_{ij} be the number of elements where both X_i and Y_j are assigned. Let $n_{i.}$ and $n_{.j}$ be the number of elements where labels X_i and Y_j are assigned, respectively. Then, ARI score of label assignments $\mathcal{X}$ and $\mathcal{Y}$ is defined as follows:

$$\mathrm{ARI}(\mathcal{X}, \mathcal{Y}) = \frac{\sum_i \sum_j \binom{n_{ij}}{2} - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{n_{i.}}{2} + \sum_j \binom{n_{.j}}{2}\right] - \left[\sum_i \binom{n_{i.}}{2} \sum_j \binom{n_{.j}}{2}\right] / \binom{n}{2}}, \tag{5}$$

where $\binom{a}{b}$ denotes the number of b-combination of a distinct elements. We used this criterion to measure similarities between the label assignments by clustering and the ground truth labels.

C. EVALUATION OF THE NUMBER OF USER FEEDBACK

Firstly, we evaluated the performance of the system when a user uses it for a year ($D = 365$ iterations). The user gives feedback whenever the clustering result is wrong. In terms of the user experience, the number of user feedback should be small. If the feature extractor could extract appropriate features, clustering results would be accurate, and the number of user feedback would be small.

TABLE 1. The average number of required user feedback in 365 iterations.

Method	# of user feedback ↓
Interactive user adaptation (Proposed)	**71.0**
Only label modification (without adaptation)	200.8

We compared the proposed method with a method that does not update the feature space (without adaptation). The method (without adaptation) only modifies wrong labels by the user feedback and never updates the feature space.

Table 1 shows the average number of user feedback in a year ($D = 365$) over the 120 sequences. It shows that the proposed interactive user adaptation method significantly reduces the number of user feedback by updating the feature extractor. The average number of required user feedback in the 365 iterations was 71.0. From the results, if a user gives only one feedback within about five days on average, the system works almost perfectly.

D. CLUSTERING PERFORMANCE AGAINST THE NUMBER OF USER FEEDBACK

We evaluated the clustering performance when the maximum number of user feedback is limited. We limited the maximum number of user feedback A to 0 (no feedback), 10, 40, and 100 within a year ($D = 365$ iterations). Then, we evaluated the clustering performance at the end of the image sequences.

We compared the performance of the proposed method to the situation with no user adaptation. As a comparative method, we used a method (without adaptation) that modifies the clustering label but does not update the feature extractor; that is, it does not adapt to the user. We also compare with a method that does not use any user feedbacks (No user feedback).

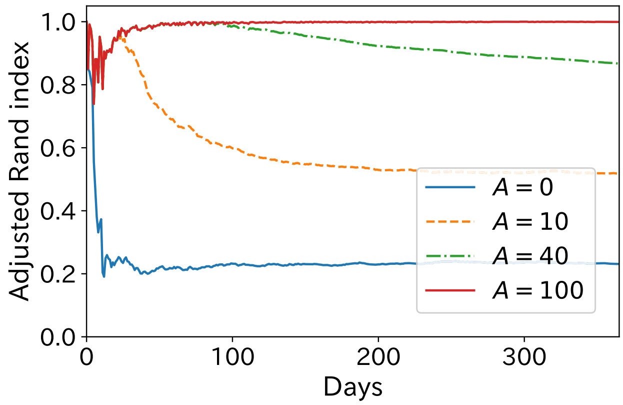
By comparing to the method with only label modification (without adaptation), we confirmed that the proposed interactive user adaptation achieved higher accuracy.

To investigate the effects of user feedback, we show the transition of the ARI scores. The results are summarized in Table 2, and visualized in Fig. 11.

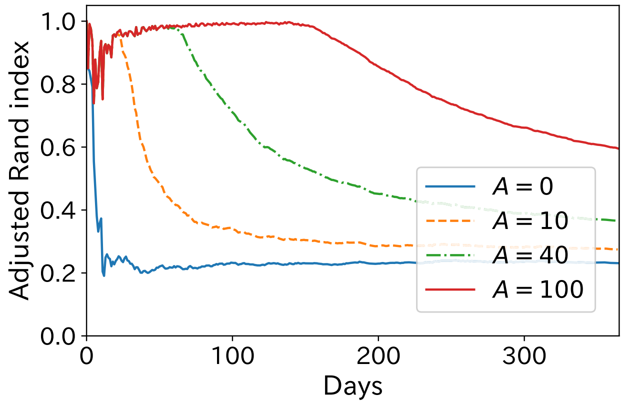
From Fig. 11, it is confirmed that more user feedback gives more accurate results. At the beginning of the steps, the ARI score shows noisy behavior because of the small number of samples. We can see that the ARI score rapidly drops when the number of feedbacks reaches the limit, in case the feature extractor is not updated. In addition, we can see that the ARI score decreases if the number of user feedback is small. However, the ARI scores decrease slowly and are high until the end of the iterations. In case the limit of the number of user feedback is 100, the number of user feedback does not reach the limit even upon 365 iterations for the proposed method.

TABLE 2. Clustering evaluation after $D = 365$ iterations.

Method	Adjusted Rand index ↑
# of maximum user feedback $A = 10$	
Interactive user adaptation (Proposed)	**0.518**
Only label modification (without adaptation)	0.275
No user feedback	0.231
# of maximum user feedback $A = 40$	
Interactive user adaptation (Proposed)	**0.867**
Only label modification (without adaptation)	0.366
No user feedback	0.231
# of maximum user feedback $A = 100$	
Interactive user adaptation (Proposed)	**0.999**
Only label modification (without adaptation)	0.596
No user feedback	0.231



(i) The proposed interactive user adaptation by user feedback.



(ii) Only label modification by user feedback.

FIGURE 11. Transition of the adjusted Rand index score for 365 iterations: the difference between these lines is the maximum number of user feedback.

E. CLUSTERING PERFORMANCE ON KNOWN DATA

We evaluated the clustering performance after the system had adapted to a user, where all the clothing instances of the user are known. First, we ran the system for 365 iterations with user feedback (without limitation on the maximum number of user feedback). The system adapted suitably to the user. Then, we additionally ran the system for 365 iterations without user feedback and evaluated the clustering performance at the last iteration. Here, in the additional 365 iterations, it is assumed that all the clothing is known; that is, there is no new clothing instance in the 365 iterations.

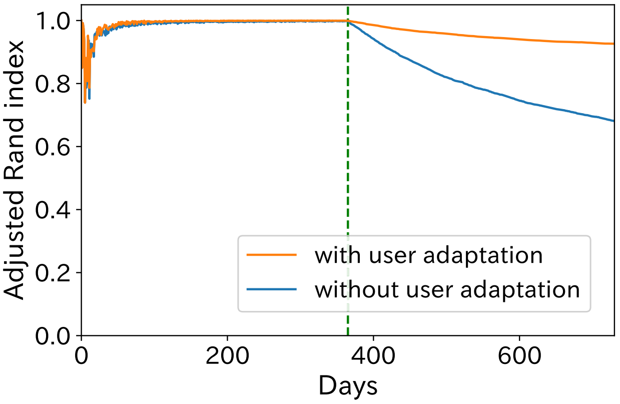


FIGURE 12. Transition of the adjusted Rand index score for $D = 730$ iterations: the vertical dashed line at the center of the graph shows the end of user feedback. On the right side of the vertical dashed line, the user did not give any feedback.

To evaluate the effectiveness of the user adaptation, we compare the two methods whose feature extractors are updated or not in the former 365 iterations. We evaluated the performance by using the ARI score.

The results are shown in Fig. 12. As shown in Fig. 12, the system with user adaptation (updating the feature extractor) exhibited high performance in the latter 365 iterations.

F. VISUALIZATION OF THE FEATURE SPACES

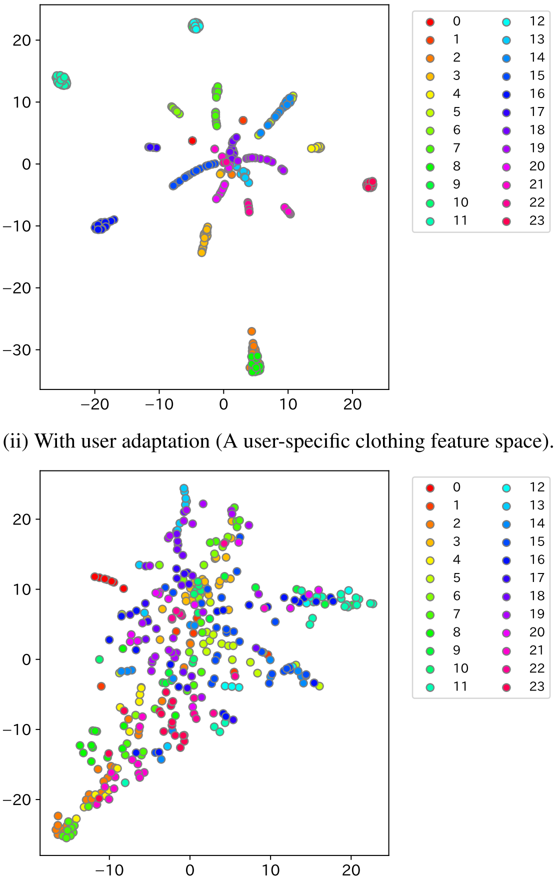
To investigate the effect of the user adaptation method, we visualized the generic and a user-specific clothing feature space. After 365 iterations of clustering with and without user adaptation, we visualize the features by applying t-distributed Stochastic Neighbor Embedding (t-SNE) [39] to obtain two-dimensional spaces. In Fig. 13, different pieces of clothing are indicated in different colors. By the proposed method (Fig. 13 (i)), the overlapping of the feature points is reduced, and features of each piece of clothing are gathered closer while the distance between each cluster is larger. It makes clustering easier and much accurate. By these observations, we confirmed that the proposed method effectively adapts the feature space to the user's clothing.

G. CLUSTERING PERFORMANCE WHEN THE USER FEEDBACK IS NOT ALWAYS CORRECT

Users are prone to making mistakes such as forgetting to provide user feedback or providing wrong feedback. In such cases, mislabeled samples may be included in the clustering results.

To evaluate the robustness against mislabeled samples, we evaluated the clustering performance by changing the ratio of correct user feedback among 0.25, 0.50, 0.75, and 1.00.

The ARI scores after 365 iterations are shown in Table 3. As a result, even if the user feedback included wrong labels, the system worked well. Even if 50% of the user feedback was incorrect, the system kept the clustering performance at about 0.7 in the ARI score. Therefore, even if the user made several



(ii) With user adaptation (A user-specific clothing feature space).

(i) Without user adaptation (The general clothing feature space).

FIGURE 13. t-SNE visualization results of clothing feature spaces of a user with and without user adaptation after 365 iterations: different colors indicate different pieces of clothing.

TABLE 3. ARI scores after 365 iterations with incomplete user feedback.

Method	Correct user feedback ratio	Adjusted Rand index ↑
Interactive user adaptation (Proposed)	(no feedback) 0	0.231
	0.25	0.554
	0.50	0.687
	0.75	0.874
	(all feedback are correct) 1	0.999

mistakes in the feedback, it is considered that the mistakes do not significantly affect performance.

VI. CONCLUSION

In this paper, we proposed a system that can automatically aggregate clothing and visualize the frequency at which a person wears them by only observing the person using a monitoring camera at home. Since there is a gap between the clothing of a person and those available in the world, it is not suitable to cluster the clothing of the user using the generic clothing features extracted by a feature extractor trained using a large number of clothing photos. To fill the gap, we proposed an interactive domain adaptation using user feedback.

Through the evaluation of a real-world dataset, we confirmed that the proposed interactive domain adaptation achieved higher clustering performance and reduced the amount of user feedback.

Creating a sophisticated user interface to make the user feedback procedure easier will be considered in a future work. In this research, we assumed that each user's clothing set is not changed during the year, including the seasonal trends of clothing. The users may dispose their clothing and buy new one. We need to consider these changes in the dataset. Since image generation techniques are actively developed recently, those methods can be used to enhance the number of clothing and people in the dataset. Evaluation on such an enhanced dataset also can be in future work.

REFERENCES

[1] Q. Tu and L. Dong, "An intelligent personalized fashion recommendation system," in *Proc. Int. Conf. Commun., Circuits Syst. (ICCCAS)*, Jul. 2010, pp. 479–485.

[2] Y. Hu, X. Yi, and L. S. Davis, "Collaborative fashion recommendation: A functional tensor factorization approach," in *Proc. 23rd ACM Int. Conf. Multimedia*, Oct. 2015, pp. 129–138.

[3] S. C. Hidayati, C.-C. Hsu, Y.-T. Chang, K.-L. Hua, J. Fu, and W.-H. Cheng, "What dress fits me best: Fashion recommendation on the clothing style for personal body shape," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 438–446.

[4] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, and H. Zha, "Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation," in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 765–774.

[5] Y.-G. Shin, Y.-J. Yeo, M.-C. Sagong, S.-W. Ji, and S.-J. Ko, "Deep fashion recommendation system with style feature decomposition," in *Proc. IEEE 9th Int. Conf. Consum. Electron. (ICCE-Berlin)*, Sep. 2019, pp. 301–305.

[6] B. Li, T. N. Sainath, A. Narayanan, J. Caroselli, M. Bacchiani, A. Misra, I. Shafran, H. Sak, G. Pundak, K. Chin, K. C. Sim, R. J. Weiss, K. W. Wilson, E. Variani, C. Kim, O. Siohan, M. Weintraub, E. McDermott, R. Rose, and M. Shannon, "Acoustic modeling for Google home," in *Proc. INTERSPEECH*, Aug. 2017, pp. 399–403.

[7] S. Applin, "Amazon's echo look: Harnessing the power of machine learning or subtle exploitation of human vulnerability?" *IEEE Consum. Electron. Mag.*, vol. 6, no. 4, pp. 125–127, Oct. 2017.

[8] M. Ackerman and S. Dasgupta, "Incremental clustering: The case for extra clusters," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2014, pp. 307–315.

[9] J. Liu and H. Lu, "Deep fashion analysis with feature map upsampling and landmark-driven attention," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, vol. 11131, Jan. 2019, pp. 30–36.

[10] X. Gu, Y. Wong, L. Shou, P. Peng, G. Chen, and M. S. Kankanhalli, "Multi-modal and multi-domain embedding learning for fashion retrieval and analysis," *IEEE Trans. Multimedia*, vol. 21, no. 6, pp. 1524–1537, Jun. 2019.

[11] S. Wazarkar and B. N. Keshavamurthy, "Social image mining for fashion analysis and forecasting," *Appl. Soft Comput.*, vol. 95, Oct. 2020, Art. no. 106517.

[12] X. Liu, J. Li, J. Wang, and Z. Liu, "MMFashion: An open-source toolbox for visual fashion analysis," May 2020, *arXiv:2005.08847*. [Online]. Available: http://arxiv.org/abs/2005.08847

[13] K. Meshkini, J. Platos, and H. Ghassemain, "An analysis of convolutional neural network for fashion images classification (fashion-MNIST)," in *Proc. 4th Int. Sci. Conf. Intell. Inf. Technol. Ind.*, Dec. 2020, pp. 85–95.

[14] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis, "Learning fashion compatibility with bidirectional LSTMs," in *Proc. 25th ACM Int. Conf. Multimedia*, Oct. 2017, pp. 1078–1086.

[15] W. Yu, H. Zhang, X. He, X. Chen, L. Xiong, and Z. Qin, "Aesthetic-based clothing recommendation," in *Proc. 27th World Wide Web Conf.*, Apr. 2018, pp. 649–658.

[16] M. Unehara, Y. Hasegawa, K. Yamada, and I. Suzuki, "Interactive apparel coordination recommendation system reflecting situation and preference," in *Proc. Joint 11th Int. Conf. Soft Comput. Intell. Syst., 21st Int. Symp. Adv. Intell. Syst. (SCIS-ISIS)*, Dec. 2020, pp. 1–4.

[17] K. Abe, T. Suzuki, S. Ueta, A. Nakamura, Y. Satoh, and H. Kataoka, "Changing fashion cultures," *Comput. Res. Repository*, vol. arXiv:1703.07920, pp. 1–9, Aug. 2017.

[18] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, Jan. 2016.

[19] R. He and J. McAuley, "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering," in *Proc. 25th Int. Conf. World Wide Web*, Apr. 2016, pp. 507–517.

[20] H. Tsujita, K. Tsukada, K. Kambara, and I. Siio, "Complete fashion coordinator: A support system for capturing and selecting daily clothes with social network," in *Proc. 9th Int. Conf. Adv. Vis. Interfaces*, May 2010, pp. 127–132.

[21] A. Sato, K. Watanabe, M. Yasumura, and J. Rekimoto, "suGATALOG: Fashion coordination system that supports users to choose everyday fashion with clothed pictures," in *Proc. 15th Int. Conf. Hum.-Comput. Interact.*, Jul. 2013, pp. 112–121.

[22] J. McAuley, C. Targett, Q. Shi, and A. van den Hengel, "Image-based recommendations on styles and substitutes," in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2015, pp. 43–52.

[23] P. Tangseng, K. Yamaguchi, and T. Okatani, "Recommending outfits from personal closet," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 269–277.

[24] S. Jiang and Y. Fu, "Fashion style generator," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 3721–3727.

[25] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, "TextureGAN: Controlling deep image synthesis with texture patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8456–8465.

[26] H. Lee and S.-G. Lee, "Fashion attributes-to-image synthesis using attention-based generative adversarial network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 462–470.

[27] K. Ak, A. Kassim, J.-H. Lim, and J. Y. Tham, "Attribute manipulation generative adversarial networks for fashion images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10540–10549.

[28] W.-L. Hsiao, I. Katsman, C.-Y. Wu, D. Parikh, and K. Grauman, "Fashion++: Minimal edits for outfit improvement," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5046–5055.

[29] L. Liu, H. Zhang, X. Xu, Z. Zhang, and S. Yan, "Collocating clothes with generative adversarial networks cosupervised by categories and attributes: A multidiscriminator framework," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3540–3554, Sep. 2020.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[32] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 1735–1742.

[33] D. Chung, K. Tahboub, and E. J. Delp, "A two stream siamese convolutional neural network for person re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1992–2000.

[34] N. Wojke and A. Bewley, "Deep cosine metric learning for person re-identification," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, Mar. 2018, pp. 748–756.

[35] C. Song, Y. Huang, W. Ouyang, and L. Wang, "Mask-guided contrastive attention model for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1179–1188.

[36] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7291–7299.

[37] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, Y. Bengio and Y. LeCun, Eds., May 2015, pp. 1–15.

[38] J. M. Santos and M. Embrechts, "On the use of the adjusted rand index as a metric for evaluating supervised classification," in *Proc. 19th Int. Conf. Artif. Neural Netw.*, Sep. 2009, pp. 175–184.

[39] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.

YASUTOMO KAWANISHI (Member, IEEE) received the B.Eng. degree in engineering and the M.Inf. and Ph.D. degrees in informatics from Kyoto University, Japan, in 2006, 2008, and 2012, respectively. He became a Postdoctoral Fellow with Kyoto University, in 2012. In 2014, he moved to Nagoya University, Japan, as a Designated Assistant Professor. In 2015, he became an Assistant Professor and a Lecturer, in 2020. Since 2021, he has been the Team Leader of the Multimodal Data Recognition Research Team, RIKEN Guardian Robot Project. His main research interests include robot vision for environmental understanding and computer vision for human understanding, especially pedestrian detection, tracking, retrieval, and recognition. He is a member of IIEEJ and IEICE. He received the Best Paper Award from SPC2009 and the Young Researcher Award from the IEEE ITS Society Nagoya Chapter.

HIROSHI MURASE (Life Fellow, IEEE) received the B.Eng., M.Eng., and Ph.D. degrees in electrical engineering from Nagoya University, Japan. In 1980, he joined Nippon Telegraph and Telephone Corporation (NTT). From 1992 to 1993, he was a Visiting Research Scientist with Columbia University, New York. He has been a Professor with Nagoya University, since 2003. His research interests include computer vision, pattern recognition, and multimedia information processing. He is a fellow of the IPSJ and the IEICE. He was awarded the IEEE CVPR Best Paper Award, in 1994, the IEEE ICRA Best Video Award, in 1996, the IEICE Achievement Award, in 2002, the IEEE Multimedia Paper Award, in 2004, and the IEICE Distinguished Achievement and Contributions Award, in 2018. He received the Medal with Purple Ribbon from the Government of Japan, in 2012.

SATOSHI KOMORITA received the B.E. and M.E. degrees from The University of Tokyo, in 2004 and 2006, respectively. He joined KDDI Corporation, in 2006, and engaged in mobile network research, IEEE standardization, and smartphone development. He is currently the Senior Manager in charge of Media Recognition Laboratory at KDDI Research Inc. His current research interests include human pose recognition and position estimation from images.

SEI NAITO (Member, IEEE) received the B.E., M.E., and Ph.D. degrees from Waseda University, in 1994, 1996, and 2006, respectively. He joined KDD Corporation (currently KDDI), in 1996, and is currently the Executive Director in charge of Media ICT Division at KDDI Research Inc. His research interests include efficient video compression, media signal processing, and cross reality (XR) space creation.