J Inf Process Syst, Vol.20, No.5, pp.675~683, October 2024
https://doi.org/10.3745/JIPS.01.0109

ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Similarity Analysis Model with 6CH ResNet Structure

JunHyeok Go and Nammee Moon*

Abstract

Large-scale waste similarity analysis is crucial for automating waste management on a large scale. It involves confirming the match between waste discharged from homes and that collected by agencies, which is essential for a stable automated system. This paper compares feature extraction methods for similarity measurement, including the scale-invariant feature transform (SIFT) algorithm with added HSV color features, convolutional neural network-based encoders, and a modified 6-channel (6CH) ResNet for end-to-end learning. The results demonstrate that the 6CH ResNet achieves up to 4.9% higher accuracy than both the basic SIFT method and encoders, as well as the SIFT algorithm with HSV color features. Implementing the 6CH ResNet in automated waste management systems can enhance object similarity measurement while using fewer computing resources.

Keywords

Convolutional Neural Network (CNN), Image Similarity, Large Waste

1. Introduction

The increase in waste generation due to urbanization has become a significant global issue. As a result, numerous studies have been conducted on automating waste management, covering domestic, medical, and food waste [1,2]. However, these studies mainly focus on classifying small or already collected waste. A major challenge in automating large waste management is ensuring the match between the large waste generated in homes and that collected by agencies. This is critical because varying charges are applied based on the waste's size and type, and discrepancies can lead to errors in the management system.

For similarity analysis, methods employing the scale-invariant feature transform (SIFT) matching algorithm [3,4] and those using encoders [5] have been explored. The SIFT algorithm, relying on keypoints and descriptors, loses color information during grayscale processing for edge detection and noise reduction. On the other hand, the encoder-based method, which uses Euclidean distance to map feature vectors of each image through encoders, suffers from computational inefficiency.

To overcome these limitations, this paper presents experiments that integrate the HSV color space features lost in the SIFT process and enhance computational efficiency and accuracy by combining images into a 6-channel (6CH) format and processing them through a single network, instead of separate encoders.

Manuscript received March 5, 2024; accepted April 18, 2024.

* **Corresponding Author:** Nammee Moon (nammee.moon@gmail.com)
School of Computer Science and Engineering, Hoseo University, Asan, Korea (sunsu970306@gmail.com, Nammee.moon@gmail.com)

2. Related Work

2.1 Feature Point Extraction Algorithm

Feature point extraction algorithms such as SIFT, ORB (Oriented FAST and Rotated BRIEF), and speeded-up robust features have been evaluated for their performance in handling image rotation, distortion, and scale changes, with SIFT demonstrating average to superior performance in these areas [6]. Given its robustness to such variations, SIFT was selected for this study, as the dataset comprised images captured from multiple angles. However, SIFT's preprocessing step converts images to grayscale for edge detection and noise reduction, resulting in the loss of color information, which is critical for distinguishing large waste items of identical shapes but different colors (Fig. 1). To address this, the current study enhances the SIFT algorithm by incorporating the histogramized HSV color space to compensate for the lost color data.

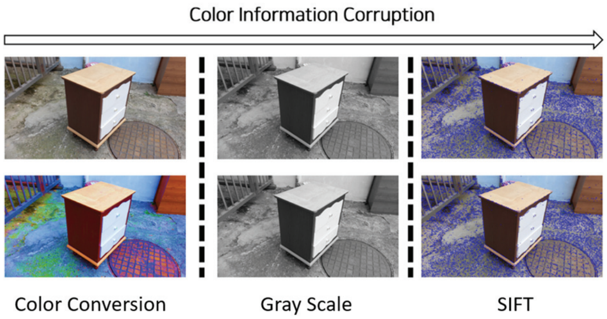

Fig. 1. Loss of color information in SIFT feature extraction.

2.2 Encoder-based Similarity Measurements

Encoders are utilized for feature extraction and dimensionality reduction across various data types, including images, text, and audio [7-9]. They effectively map high-dimensional features to a lower-dimensional space while preserving essential data characteristics, thus capturing more general features. This attribute facilitates the learning of generalized features during training, improving computational efficiency. In the context of similarity measurement, encoders are primarily employed to map data to a lower-dimensional space, simplifying the comparison of generalized features. Through this process, similar images are mapped to proximate points in the reduced space, while dissimilar ones are positioned further apart. In this study, waste images are processed through an encoder, and similarity is assessed by measuring the distances between the resulting low-dimensional representations

3. Dataset

The dataset comprises domestic waste images provided by AI Hub, captured from various angles to facilitate the training of object similarity algorithms. As illustrated in Fig. 2, the dataset construction involved pairing 8 images, 4 from each of two distinct objects, to form 8 pairs—4 of identical objects and 4 of different ones, ensuring no image reuse within the same pair category. Given the greater variety of distinct object images, numerous pairing combinations were possible. To maintain dataset balance and avoid bias, an equal number of pairs were created for both categories, with careful matching to ensure consistency in waste type.

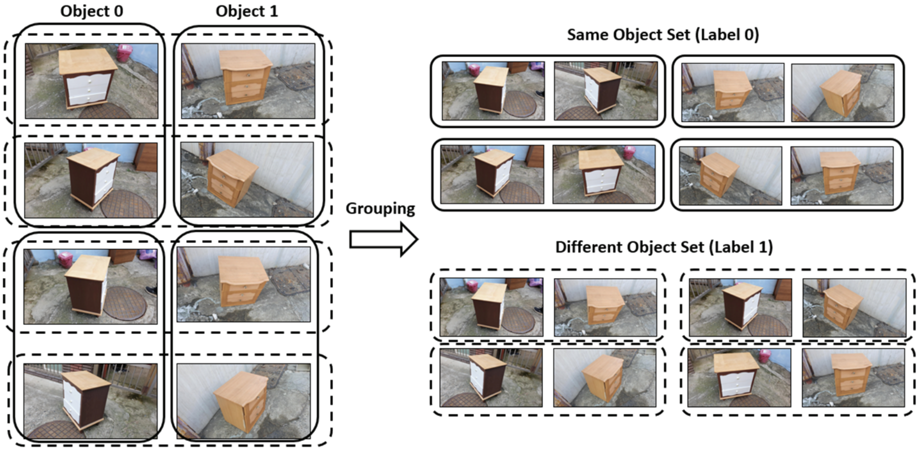

Fig. 2. Example of creating a dataset used for learning.

The dataset ultimately consisted of 110,698 pairs, with an equal distribution of 55,349 pairs each for identical and different objects, derived from 110,698 images. Table 1 details the types and quantities of large waste items used in training, selected from actual household waste. Items not categorized as large waste, such as PET bottles, glass bottles, and plastics, were excluded from the dataset. All experiments adhered to a training, validation, and test data ratio of 6:3:1, with a consistent image resolution of 512 pixels for all experiments.

Table 1. Types and number of large wastes used in experiments

Object name	Count	Object name	Count
Electric Heater	764	Fry fan	4,208
Flowerpot	6,688	Toy	4,968
Bicycle	27,944	Chest	3,944
Chair	13,696	Computer	2,000
ttukbaegi	1,144	Jar	1,400
Frame	2,096	Desk	1,368
Rice cooker	2,144	Three-wheeled bicycle	448
Table	5,080	Air purifier	300
Microwave	440	Copy machine	240
Wardrobe	430	Electric Frypan	160
Audio	1,064	Four wheeled Bicycle	2,848
Fan	2,116	Printer	1,750
Closet	4,960	Kettle	408
Iron	710	Vacuum cleaner	2,368
Oven	2,116	Mixer	716
Golf Club	4,480	TV	1,310
Air conditioner	144	Chopping board	1,864
Sofa	1,264	nightstand	800
Water purifier	264	Refrigerator	976
Video Player	630	Washing Machine	360

4. Experiment

4.1 SIFT Similarity Measurement with Added HSV

The process of measuring similarity using SIFT in conjunction with HSV involves extracting keypoints and color features through SIFT feature point extraction and HSV histogram analysis from two images. The keypoints identified by SIFT are matched using the Euclidean distance, while the similarity of color histograms is assessed using the Bhattacharyya distance for the HSV color space. These similarity metrics serve as inputs for binary classification using decision tree and support vector machine (SVM) algorithms, with the accuracy of each method being evaluated. The procedure is illustrated in Fig. 3.

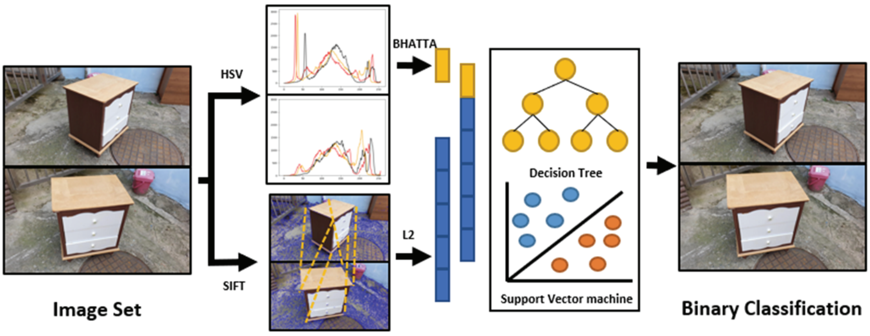

Fig. 3. Process of SIFT similarity measurement with added HSV.

In the SIFT algorithm, feature points are quantified using the $L2_{Norm}$ (Eq. 1), which is the sum of the squared differences between two vectors A and B across each dimension. Counts of matched feature points at feature distance ratios of 40%, 50%, 60%, and 70% are utilized as features. HSV histogram similarity is determined using the Bhattacharyya distance (Eq. 2), where H_1 and H_2 represent two probability distributions, and I denotes an interval. Given that SIFT-derived features are natural numbers greater than zero and HSV features calculated by Bhattacharyya distance are real numbers between 0 and 1, both standardization and normalization were applied to scale the input data to a mean of 0 and a variance of 1. This scaling is crucial to prevent learning issues caused by differences in scale. Table 2 presents the covariance, correlation coefficient (Eq. 3), and p-value results, assessing the correlation of the extracted HSV features in large waste similarity measurement. The analysis reveals that HSV features are positively correlated with similarity, whereas SIFT features show a negative correlation, indicating significant correlations for both.

$$L2_{Norm} = \sum_i (A_i - B_i)^2, \tag{1}$$

$$BATTA(H_{1,}H_2) = \sqrt{1 - \frac{1}{\sqrt{H_1 H_2 N^2}} \sum_i \sqrt{H_1(I) \cdot H_2(I)},} \tag{2}$$

$$\mathrm{Cov(x,y)} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n}, \mathrm{Corr(x,y)} = \frac{Cov(x,y)}{\sigma x \cdot \sigma y}. \tag{3}$$

The performance outcomes derived from utilizing SIFT and HSV features were obtained through training decision tree and SVM models, with parameters optimized based on the training set. The decision tree parameters included a Gini criterion, a maximum depth of 9, a minimum samples split of 5, a minimum samples leaf of 5, maximum features set to sqrt, class weight balanced, and the best splitter. For the SVM, the cost was set at 0.1, the kernel was linear, and the image size was consistently set at 512 pixels. Table 3 showcases the training results, highlighting a significant performance boost when combining SIFT and HSV features, irrespective of the model used. The inclusion of both SIFT and HSV features showed a negligible performance disparity between Decision Tree and SVM, with a difference of less than 0.4%. The use of color information alone in large waste resulted in a performance enhancement of 10.1% for the Decision Tree and 11.9% for the SVM, as measured by accuracy.

Table 2. Covariance, correlation coefficient, *p*-value of extracted SIFT, HSV features

	HSV	SIFT			
		40%	50%	60%	70%
Covariance	0.089	−4.516	−3.737	−4.142	−4.293
Correlation	0.788	−0.202	−0.287	−0.336	−0.369
p-value	0.0	0.0	0.0	0.0	0.0

Table 3. Result of experiments between decision tree and SVM

Model	Train feature	Accuracy	Recall	Precision	F1_score
Decision tree	SIFT	81.8	89.7	77.4	83.1
	SIFT + HSV	**91.9**	**93.8**	**90.3**	**92.0**
SVM	SIFT	79.7	95.0	72.6	82.3
	SIFT + HSV	**91.6**	**94.1**	**89.5**	**91.8**

The bold font indicate the best performance results for each test.

4.2 Measurement Similarity using Pre-trained Network

To further assess the similarity of large waste items, a method involving the extraction of feature vectors from a pre-trained convolutional neural network-based network, specifically ResNet-50, was employed. ResNet-50, which was pre-trained on the ImageNet dataset, served as the encoder. The fully connected (FC) layer of ResNet-50 was removed, allowing for the extraction of (1 × 2048) feature vectors from the last BottleNeck output for each (512 × 512 × 3) image vector, as depicted in Fig. 4. The similarity between the feature vectors of each image was quantified using the $L2_{Norm}$ (Eq. 1). As illustrated in Fig. 5, the analysis revealed that the smaller the distance between feature vectors, the closer the value is to 1, indicating higher similarity. Conversely, larger distances yield values closer to 0, denoting dissimilarity. This principle underpinned the binary classification, where a threshold was set to differentiate between similar and dissimilar images, and the performance of this method was subsequently evaluated.

The effectiveness of this approach was determined using the test set, where threshold values of 0.64 and 0.63, identified as optimal based on accuracy from the training set using the $L2_{Norm}$, were applied to evaluate the performance on the validation set. At the 0.64 threshold, the method achieved an accuracy rate of 89.8%, demonstrating its robustness in distinguishing between similar and dissimilar large waste items. The detailed experimental results are shown in Table 4.

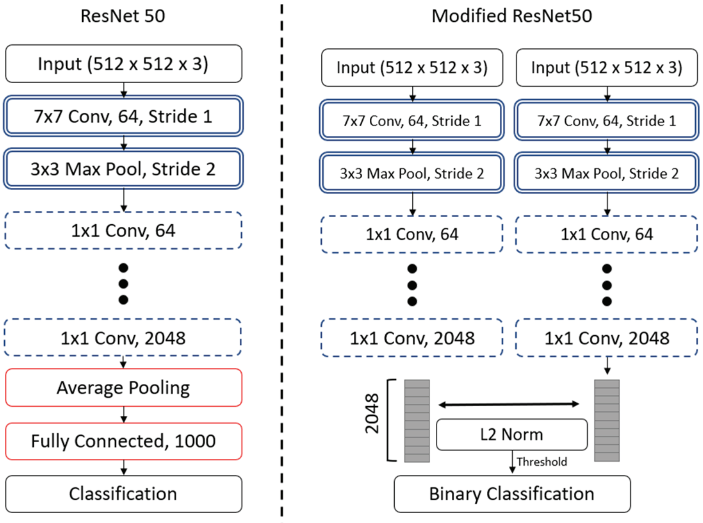

Fig. 4. Process of extracting feature vectors using a pre-trained network.

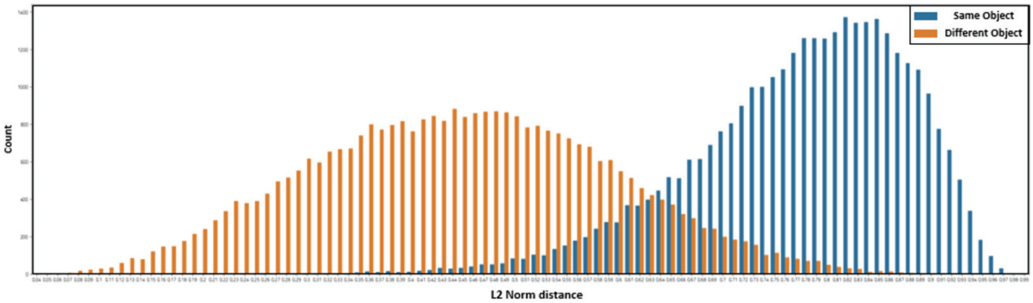

Fig. 5. Euclidean distance between feature vectors extracted using ResNet from the training set.

Table 4. Experimental results of validation data according to the $L2_{Norm}$ of image pairs

Model	$L2_{Norm}$ threshold	Accuracy	Recall	Precision	F1_score
Encoder ResNet	0.64	89.8	89.8	89.7	89.8
	0.63	89.7	88.6	90.7	89.6

4.3 Measurement of Similarity using 6CH ResNet Structure

To overcome the limitations of using ResNet as an encoder, which requires the additional computation of Euclidean distances for the feature vectors extracted from images and results in decreased computational efficiency and lower performance compared to the SIFT and HSV combination method, this study introduced a novel approach. By merging images to form a 6CH input and extracting features within a single network, the system efficiently learned common features. As depicted in Fig. 6, the 3-channel (RGB) representations of two images intended for similarity measurement were concatenated to create a 6CH input. The ResNet input layer was modified to accommodate 6CH data, and a sigmoid function was added after the FC layer for binary classification. The model was configured to predict pairs of different objects for output values above 0.5 and pairs of the same object for values below.

The training hyperparameters were set as follows: 40 epochs, a batch size of 32, a learning rate of

0.001, 8 workers, Adam optimizer, StepLR scheduler, and binary cross-entropy loss function. The experiments were conducted using two RTX 3090 GPUs, on a Linux 20.04 LTS operating system, with the training duration being approximately 22 hours. Fig. 7 illustrates the training loss graph of the 6CH ResNet structure, showing convergence of the training and validation loss values over 40 epochs.

Table 5 reveals that the 6CH ResNet method, which combines image channels to form a 6CH input, exhibited the highest performance. This marked a significant improvement of 4.9% over the least effective encoder-based ResNet method and a 2.8% increase in accuracy compared to the decision tree and SVM models trained with combined SIFT and HSV features. This demonstrates the 6CH ResNet structure's effectiveness in learning the common features of image objects. Although fine-tuning of the pre-trained ResNet was explored, it did not yield a substantial performance difference, varying only by 0.1 to 0.2.

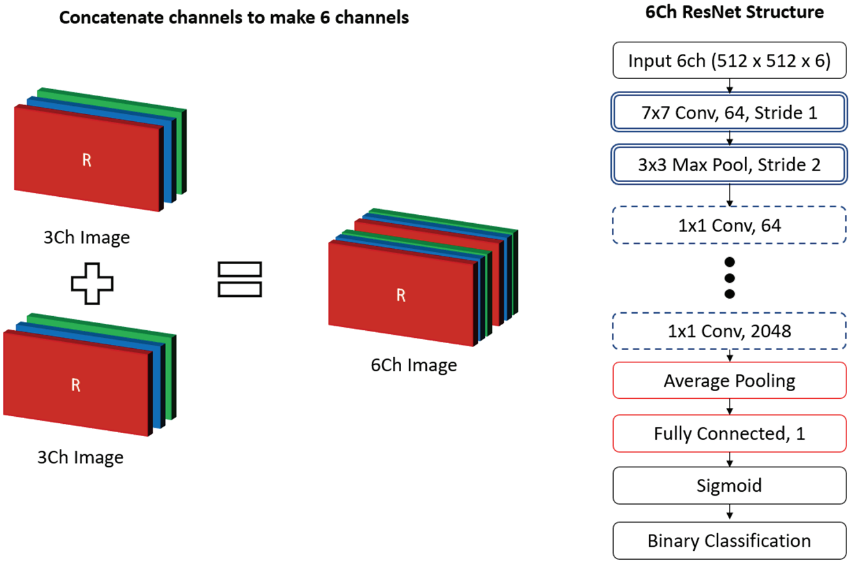

Fig. 6. Process of image concatenation and the 6CH ResNet structure.

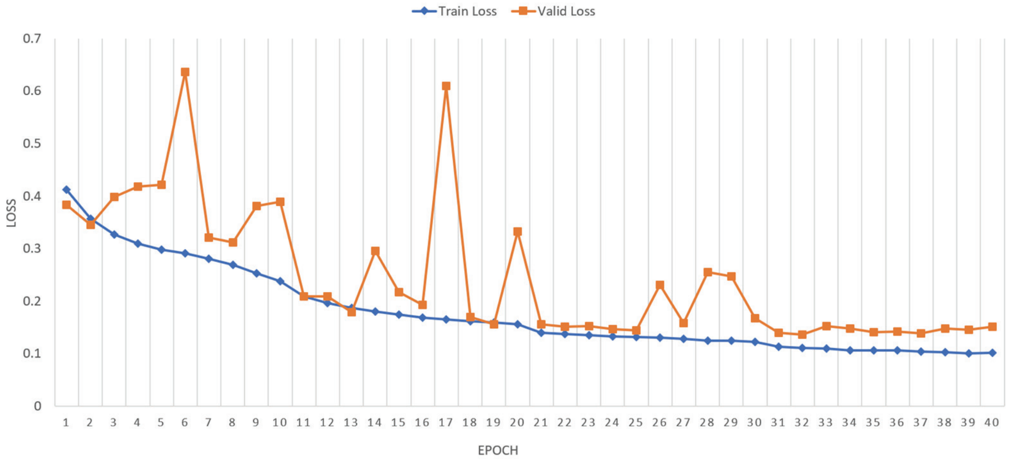

Fig. 7. Training loss graph of the 6CH ResNet structure.

Table 5. Comparing the performance evaluation of 6CH ResNet

Model	Threshold	Accuracy	Recall	Precision	F1_score
6CH ResNet	0.5	**94.7**	**94.0**	**95.0**	**94.8**
Encoder ResNet	0.64	89.7	88.6	90.7	89.8
Decision tree	SIFT + HSV	91.9	93.8	90.3	92.0
SVM	SIFT + HSV	91.6	94.1	89.5	91.8

The bold font indicate the best performance results for each test.

5. Conclusion

This study conducted extensive experiments on similarity models to enhance the automation of large waste management systems. By combining SIFT matching and HSV histogram features, an accuracy improvement of 11.9% was achieved over methods relying solely on SIFT matching. Additionally, by adjusting the input layer of ResNet to process 6CH images, a further 4.9% improvement in performance was attained. The development of an automated large waste management system incorporating a 6CH configuration is anticipated to enhance communication between households and waste collection agencies, thereby improving operational efficiency. Future work may include expanding the dataset for more comprehensive testing or employing a more compact network structure than ResNet for the 6CH configuration to further reduce computational demands.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

This paper was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1A2C2011966).

References

[1] V. Muneeswaran, P. Nagaraj, A. Akhila, L. Sudeepthi, B. Venkateswararao, and B. V. Krishna, “Smart segregation of waste and automatic monitoring system,” in *Proceedings of 2023 International Conference on Computer Communication and Informatics (ICCCI)*, Coimbatore, India, 2023, pp. 1-6. https://doi.org/10.1109/ICCCI56745.2023.10128456

[2] T. Ji, H. Fang, R. Zhang, J. Yang, L. Fan, and J. Li, “Automatic sorting of low-value recyclable waste: a comparative experimental study,” *Clean Technologies and Environmental Policy*, vol. 25, no. 3, pp. 949-961, 2023. https://doi.org/10.1007/s10098-022-02418-7

[3] L. Marlinda, F. Budiman, R. S. Basuki, and A. Z. Fanani, “Comparison of sift and orb methods in identifying the face of Buddha statue,” *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 8, no. 2, pp. 145-150, 2023. https://doi.org/10.33480/jitk.v8i2.4086

[4] A. Higaki, T. Kurokawa, T. Kazatani, S. Kido, T. Aono, K. Matsuda, et al., “Image similarity-based cardiac rhythm device identification from X-rays using feature point matching,” *Pacing and Clinical Electrophysiology*, vol. 44, no. 4, pp. 633-640, 2021. https://doi.org/10.1111/pace.14209

[5] M. Liang, R. W. Liu, S. Li, Z. Xiao, X. Liu, and F. Lu, “An unsupervised learning method with convolutional auto-encoder for vessel trajectory similarity computation,” *Ocean Engineering*, vol. 225, article no. 108803, 2021. https://doi.org/10.1016/j.oceaneng.2021.108803

[6] E. Karami, S. Prasad, and M. Shehata, “Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images,” 2017 [Online]. Available: https://arxiv.org/abs/1710.02726.

[7] K. Rama, P. Kumar, and B. Bhasker, “Deep autoencoders for feature learning with embeddings for recommendations: a novel recommender system solution,” *Neural Computing and Applications*, vol. 33, no. 21, pp. 14167-14177, 2021. https://doi.org/10.1007/s00521-021-06065-9

[8] P. Y. Huang, H. Xu, J. Li, A. Baevski, M. Auli, W. Galuba, F. Metze, and C. Feichtenhofer, “Masked autoencoders that listen,” 2022 [Online]. Available: https://arxiv.org/abs/2207.06405v1.

[9] A. Gogna and A. Majumdar, “Discriminative autoencoder for feature extraction: application to character recognition,” *Neural Processing Letters*, vol. 49, pp. 1723-1735, 2019. https://doi.org/10.1007/s11063-018-9894-5

JunHyeok Go **https://orcid.org/0000-0003-4254-5212**

He received B.S. degrees in School of Computer Science and Engineering from Hoseo University in 2023. Since March 2023, he is current with the Department of Computer Science and Engineering from Hoseo University as Master Course. His research interests include computer vision, time-series data and big data processing and analysis.

Nammee Moon **https://orcid.org/0000-0003-2229-4217**

She received B.S., M.S., and Ph.D. degrees from the School of Computer Science and Engineering at Ewha Womans University in 1985, 1987, and 1998, respectively. She served as an assistant professor at Ewha Womans University from 1999 to 2003, a then as a professor of digital media, Graduate School of Seoul Venture Information, from 2003 to 2008. Since 2008, has been a professor of computer information at Hoseo University. Her current research interests include social learning, HCI and User-centric data, and big data processing and analysis.