Deduplication of Retrieved Image Data Using Deep Network Features

School of Computer Science and Engineering
Kyoungpook National University

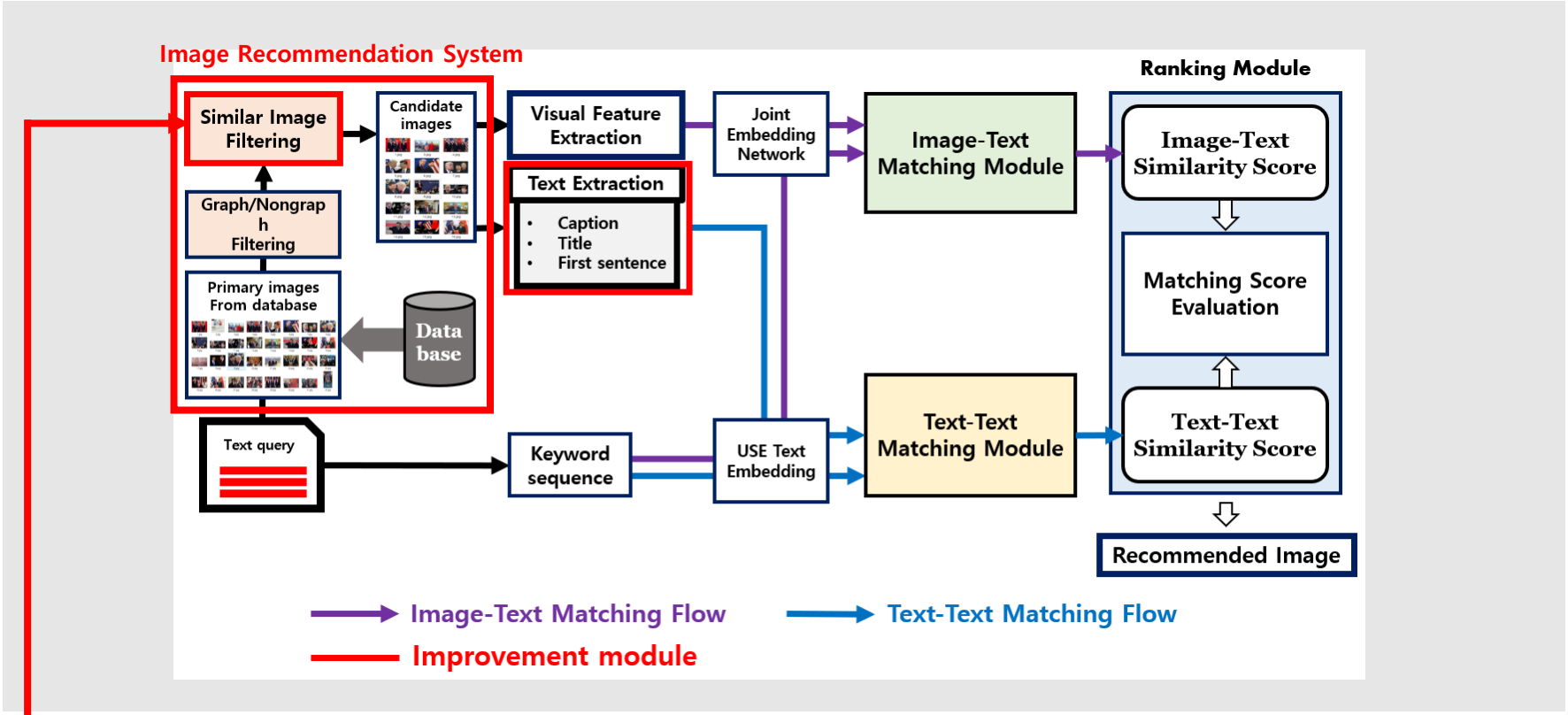
Heesung Yang
Chan Hur
Changhun Hyun
Hyeyoung Park

BCMI Lab

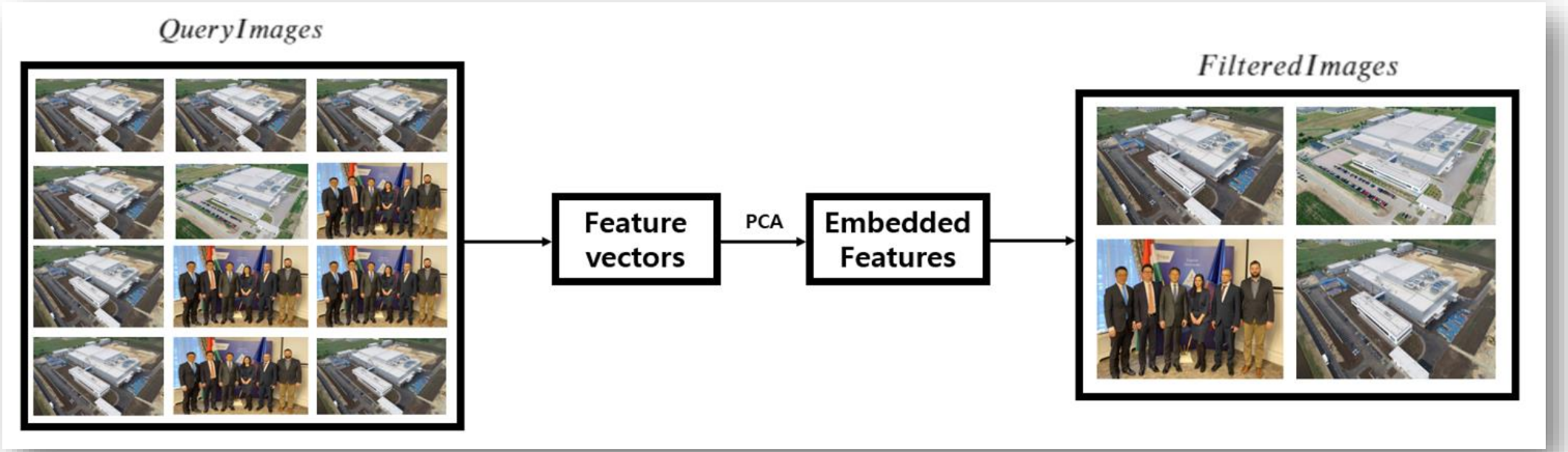
Contents

I. Abstraction

Smart Summary Report Generation System

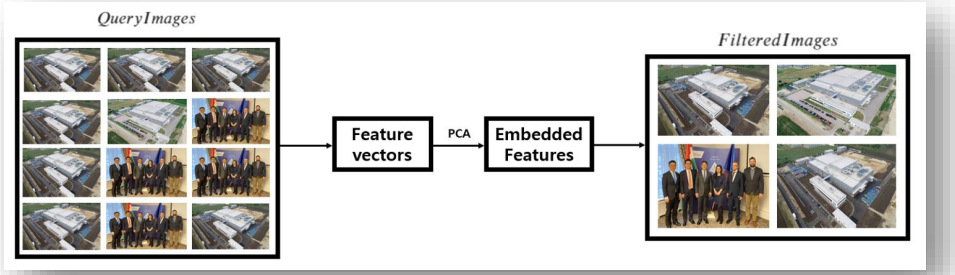


Overall Structure

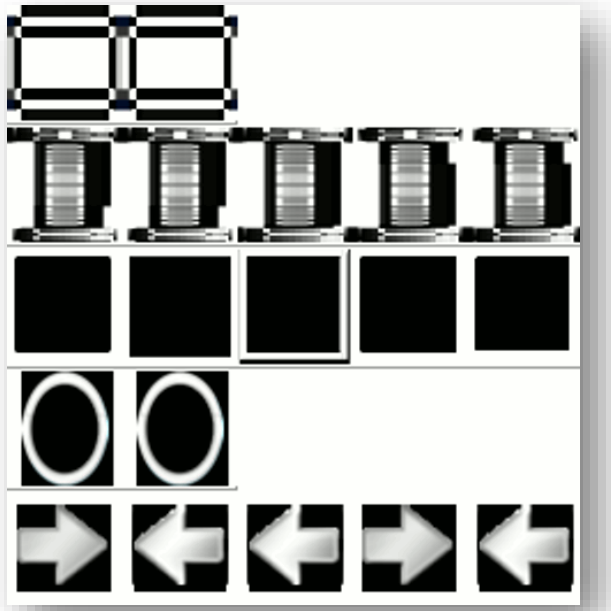


Proposed Method and VisiPics SW

dupimage(Proposed method)



VisiPics(Opensource SW)



visipics.info/

II. Introduction

Example of Duplicated Image Dataset

III. Proposed Method

Overall Process

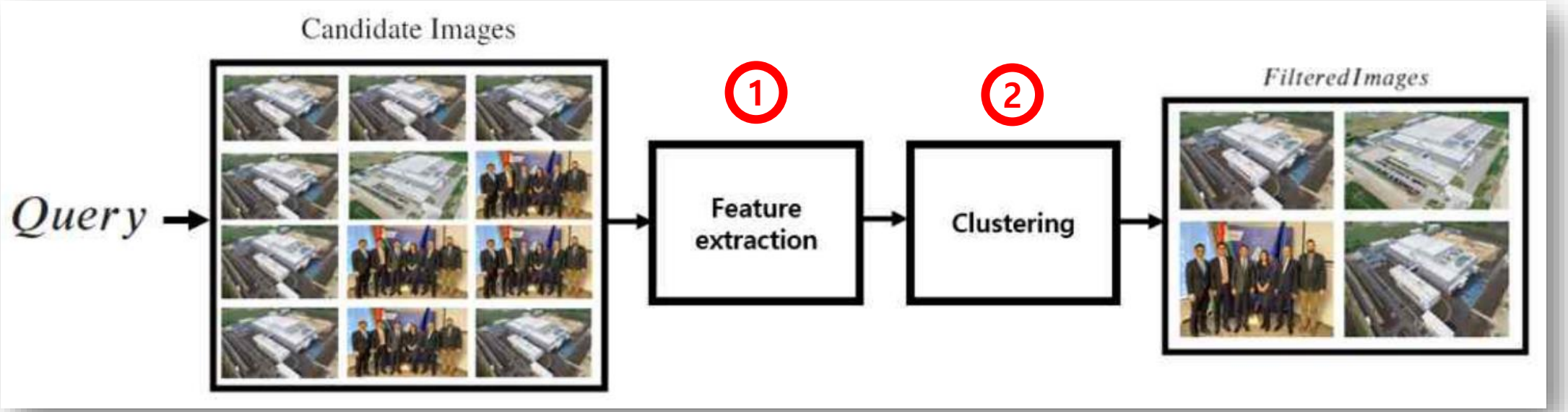


Figure 1. Overall process of the proposed module

Feature Extraction Module

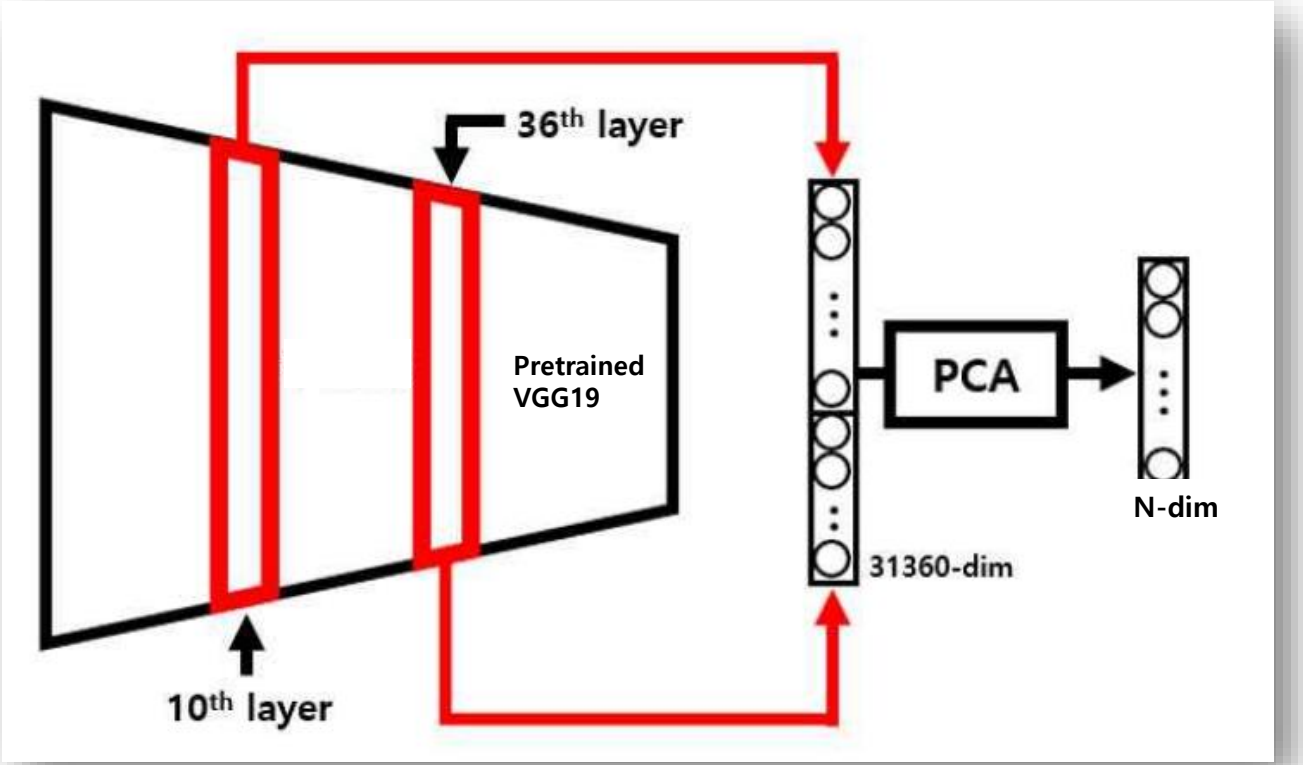
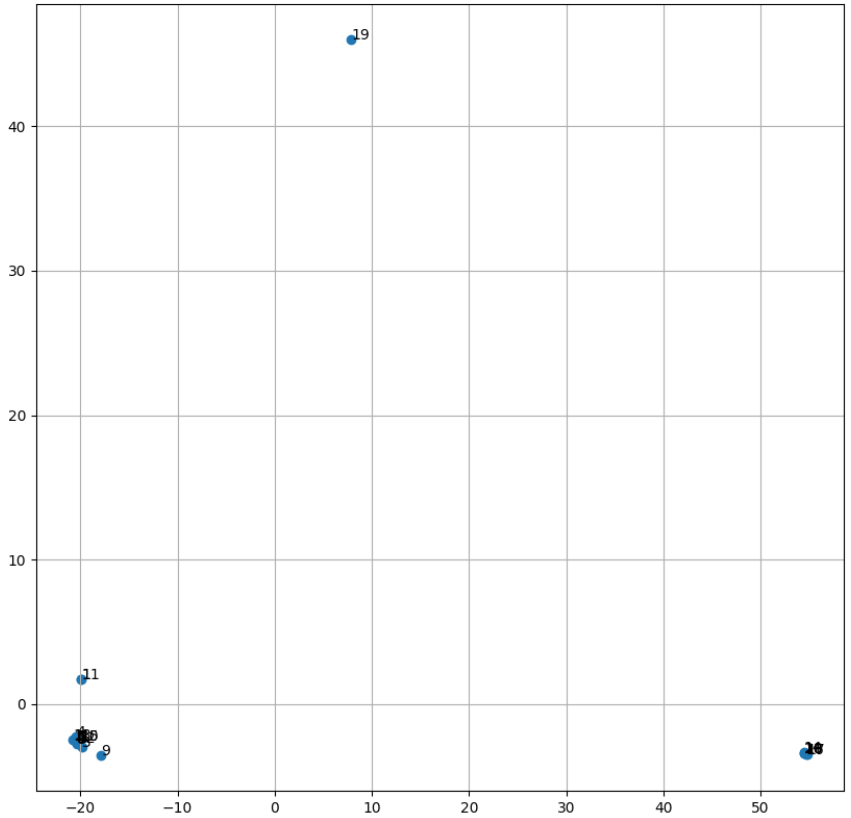
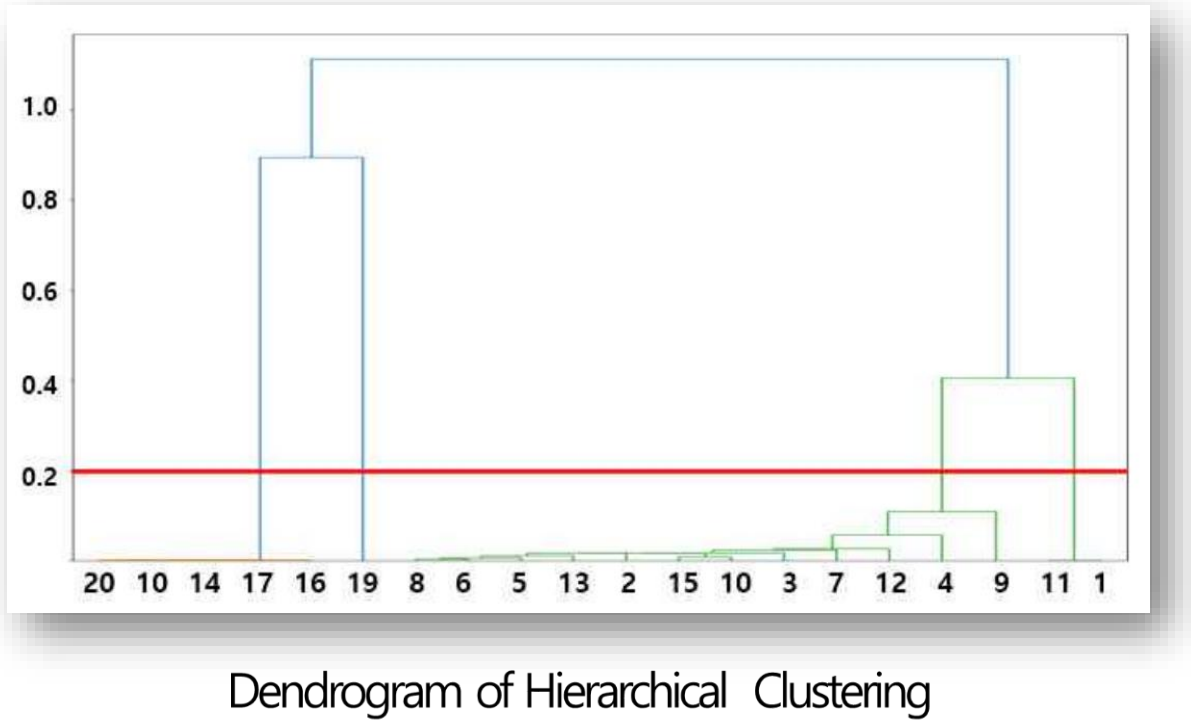


Figure 2. Feature extraction module

Extracted Features and Clustering Result



Extracted Features Plotting with PCA



Dendrogram of Hierarchical Clustering

IV. Experimental Results

Dataset

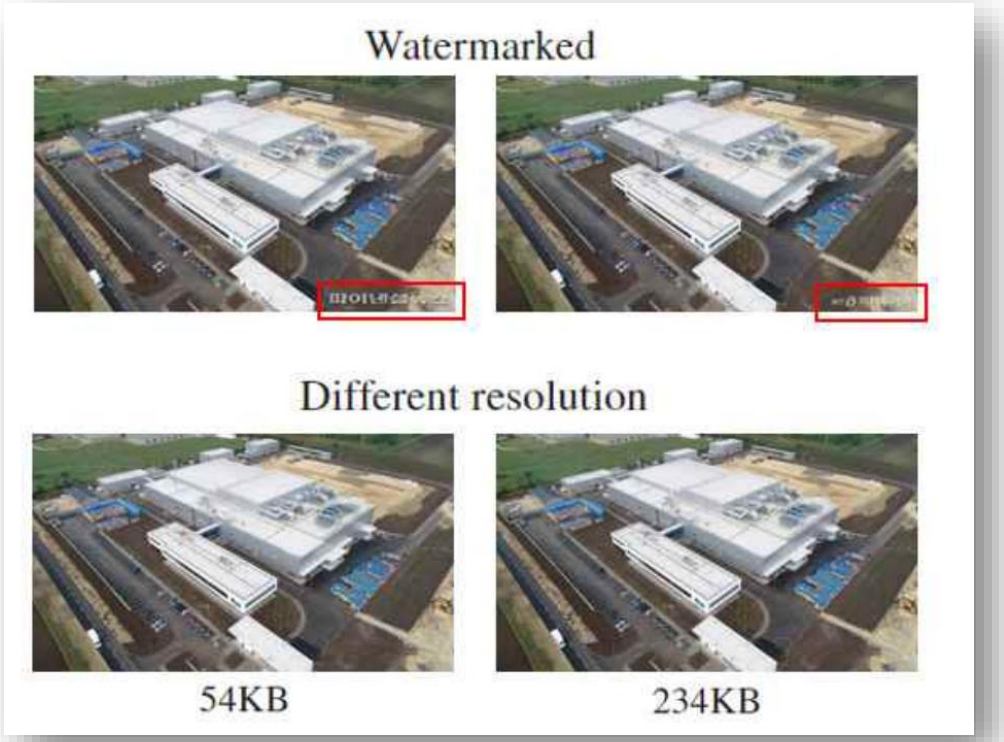
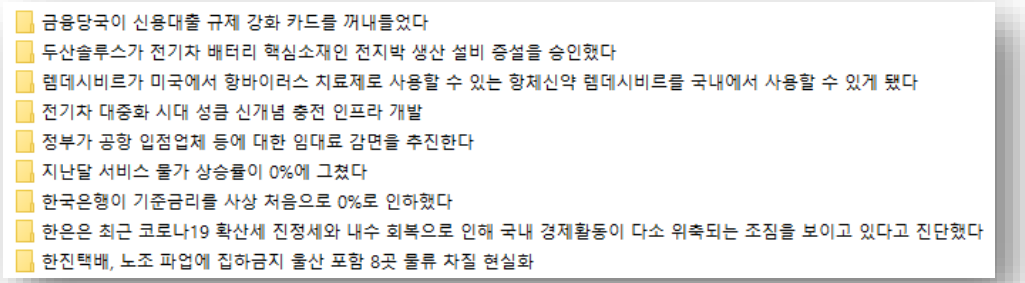


Figure 3. Example of Duplicated images in the dataset



9 Queries
20 Images per query

Evaluation Criteria

$$Precision = \frac{TP}{TP + FP},$$

$$Recall = \frac{TP}{TP + FN},$$

$$\text{F1-score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

TP is the number of pairs that are clustered together in the predicted and the ground truth partition.

TP is the number of pairs that aren't clustered together in the predicted and the ground truth partition.

FP is the number of pairs that are clustered together in the predicted but not in the ground truth partition.

FN is the number of pairs that are clustered together in the predicted but not in the ground truth partition.

Experimental Results

Method	Precision	Recall	F1-score
Raw images	0.055	0.048	0.052
Raw images + PCA	0.072	0.840	0.101
VGG features	0.062	0.144	0.070
VGG+PCA(Proposed)	1.0	0.866	0.907
VisiPics [3]	1.0	0.892	0.922

Table 1. Average performance of 9 queries

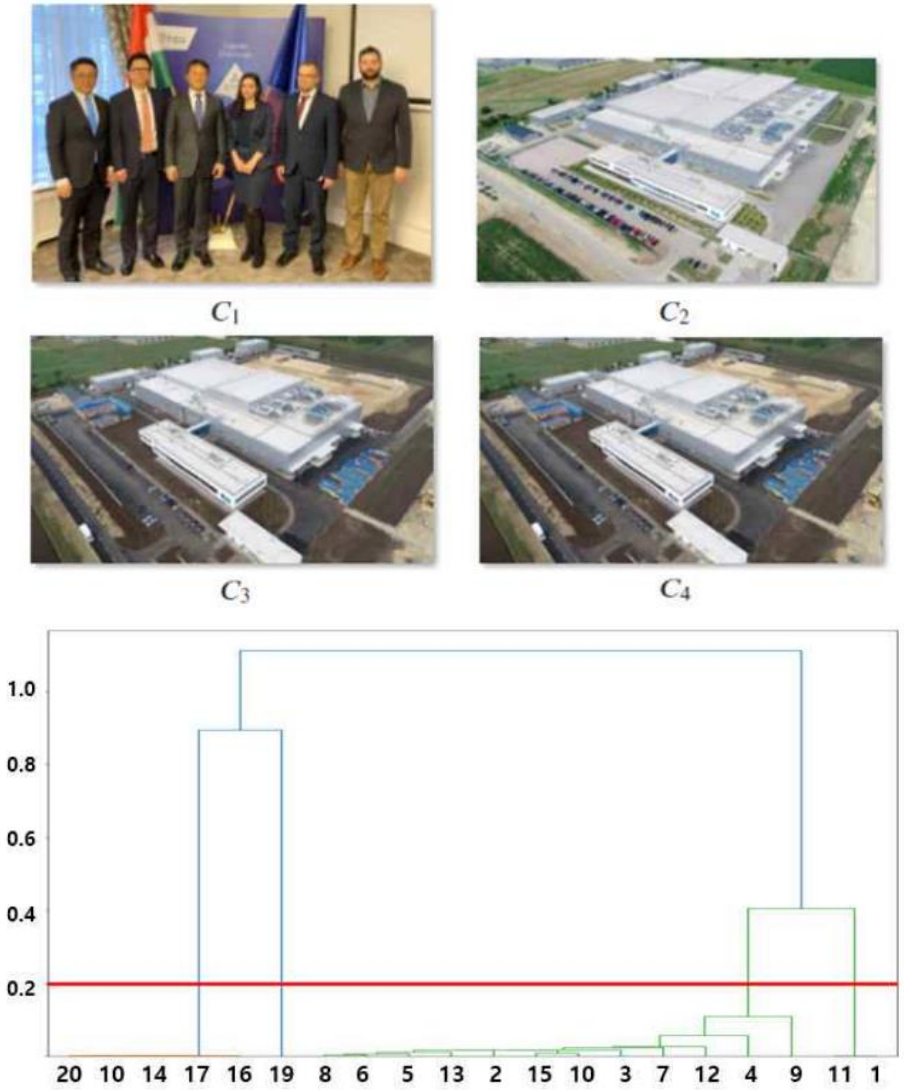
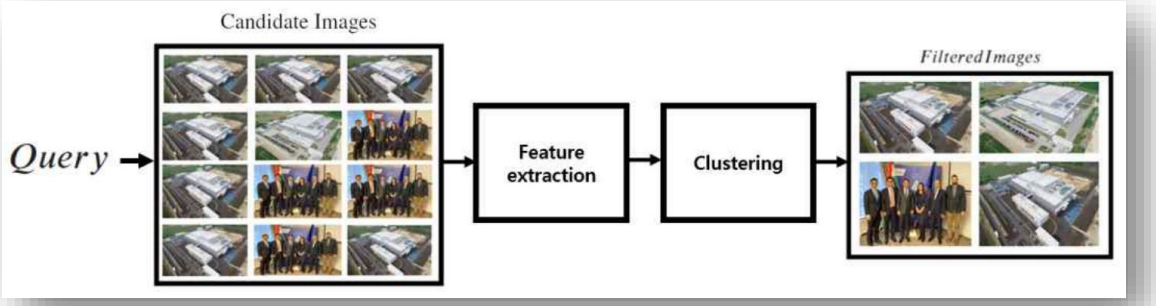
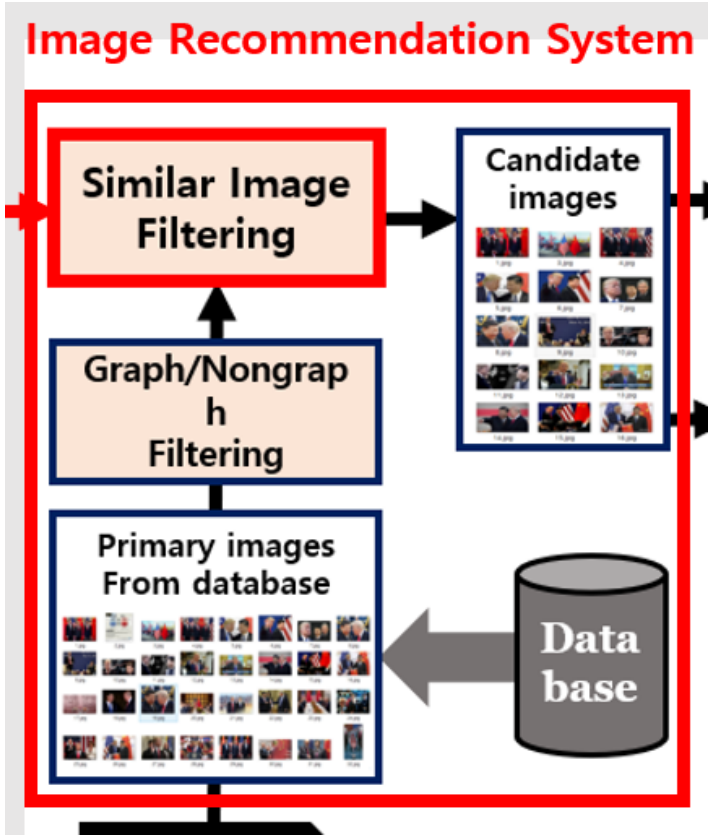


Figure 5. Result of second query

V. Conclusion

Conclusion





Reusable Pre-trained Model

VI. References

[1] John F, Gantz. "The diverse and exploding digital universe-An updated forecast of worldwide information growth through 2011." An IDC White Paper sponsored by EMC, 2008.

[2] Young Chan Moon, et al. "Data deduplication using dynamic chunking algorithm." International Conference on Computational Collective Intelligence. Springer, Berlin, Heidelberg, 2012.

[3] http://www.visipics.info, 2021.

[4] Karen Simonyan, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556, 2014.

[5] Chan Hur, Changhun Hyun, and Hyeyoung Park. "Automatic Image Recommendation for Economic Topics using Visual and Semantic Information,"2020 IEEE 14th International Conference on Semantic Computing (ICSC). IEEE, 2020.

KNU
BCMI LAB