

# **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**

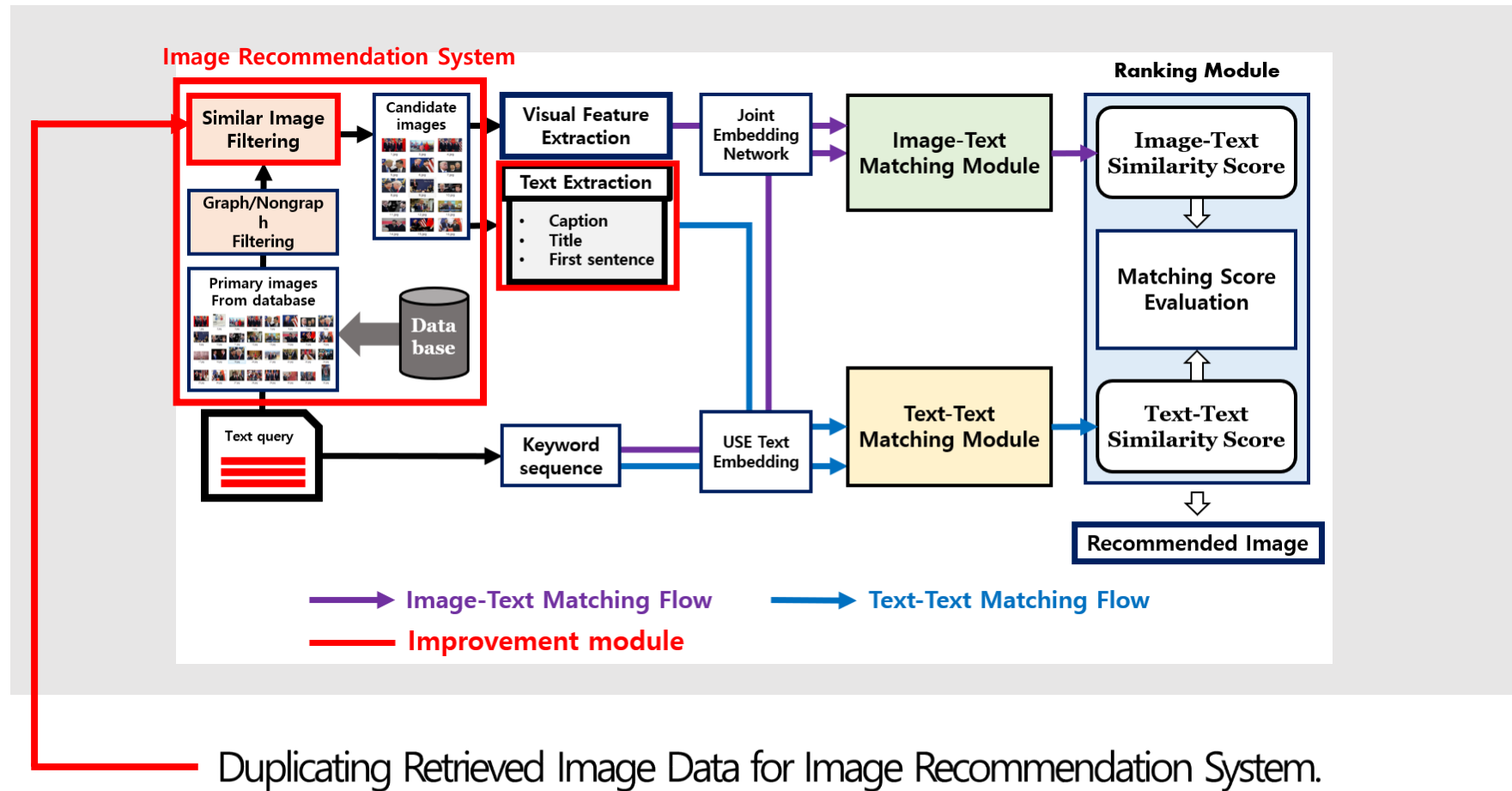


# **Contents**

- I. Abstraction**
- II. Introduction**
- III. Proposed method**
- IV. Experimental results**
- V. Conclusion**
- VI. References**

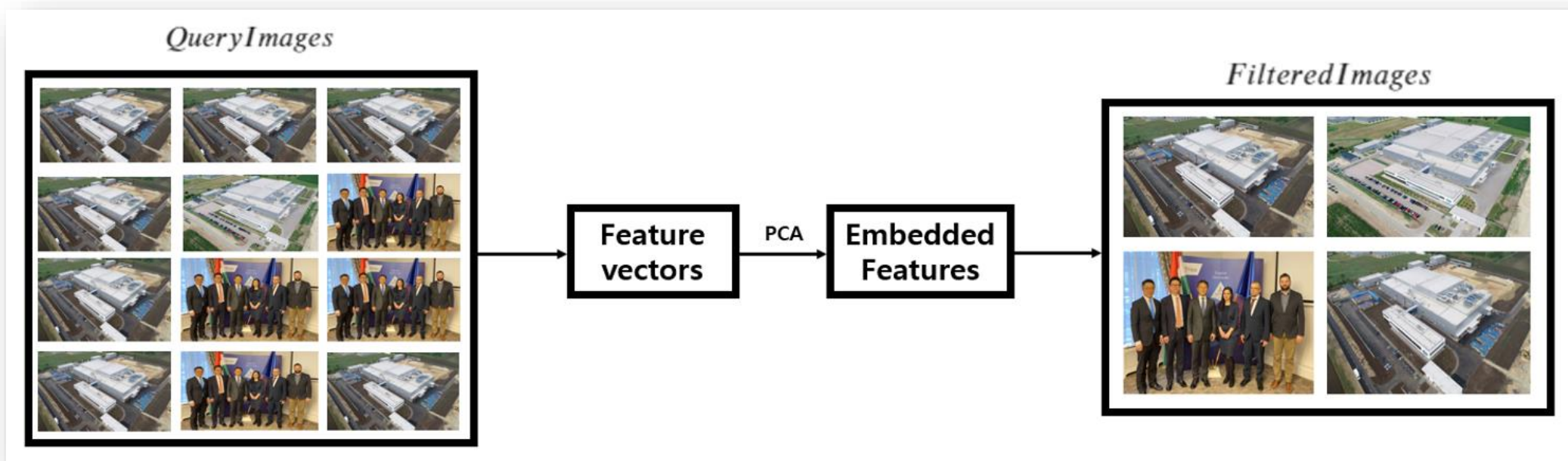
# I. Abstraction

# Smart Summary Report Generation System



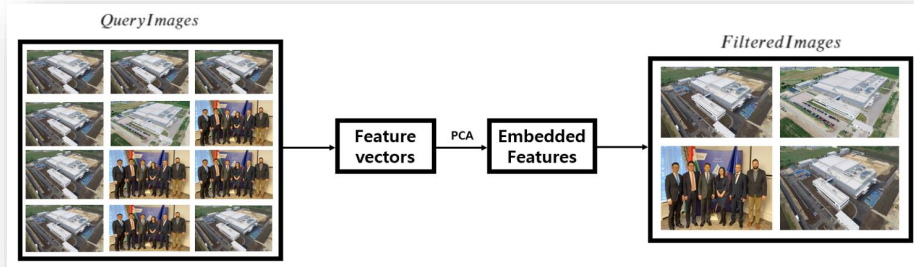


## Overall Structure

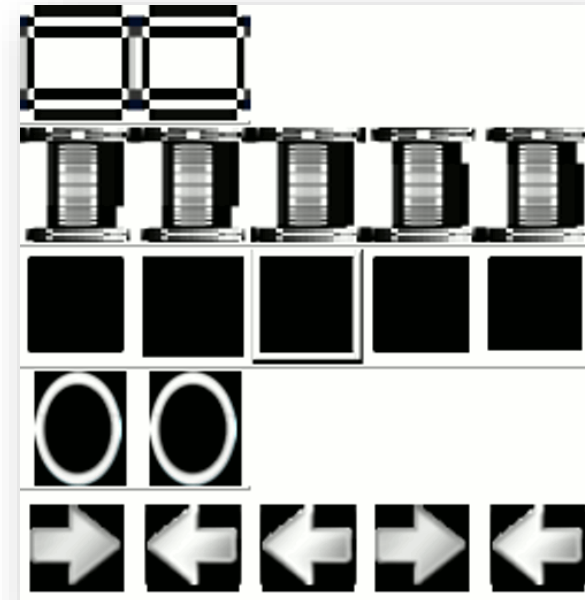


## Proposed Method and VisiPics SW

dupimage(Proposed method)



VisiPics(Opensource SW)



[visipics.info/](http://visipics.info/)

## II. Introduction

## Example of Duplicated Image Dataset





# III. Proposed Method

## Overall Process

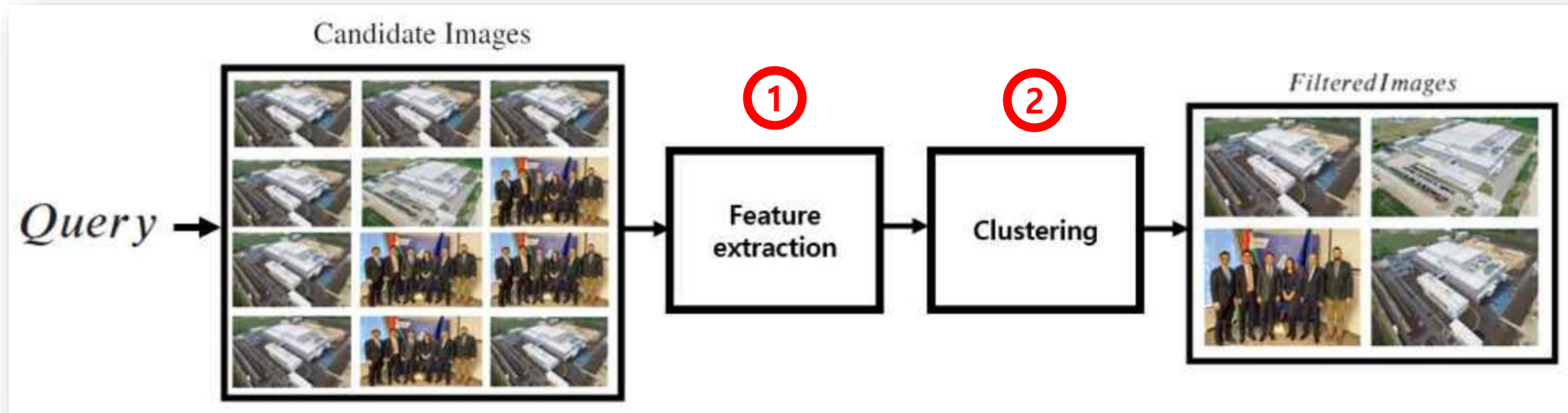


Figure 1. Overall process of the proposed module

## Feature Extraction Module

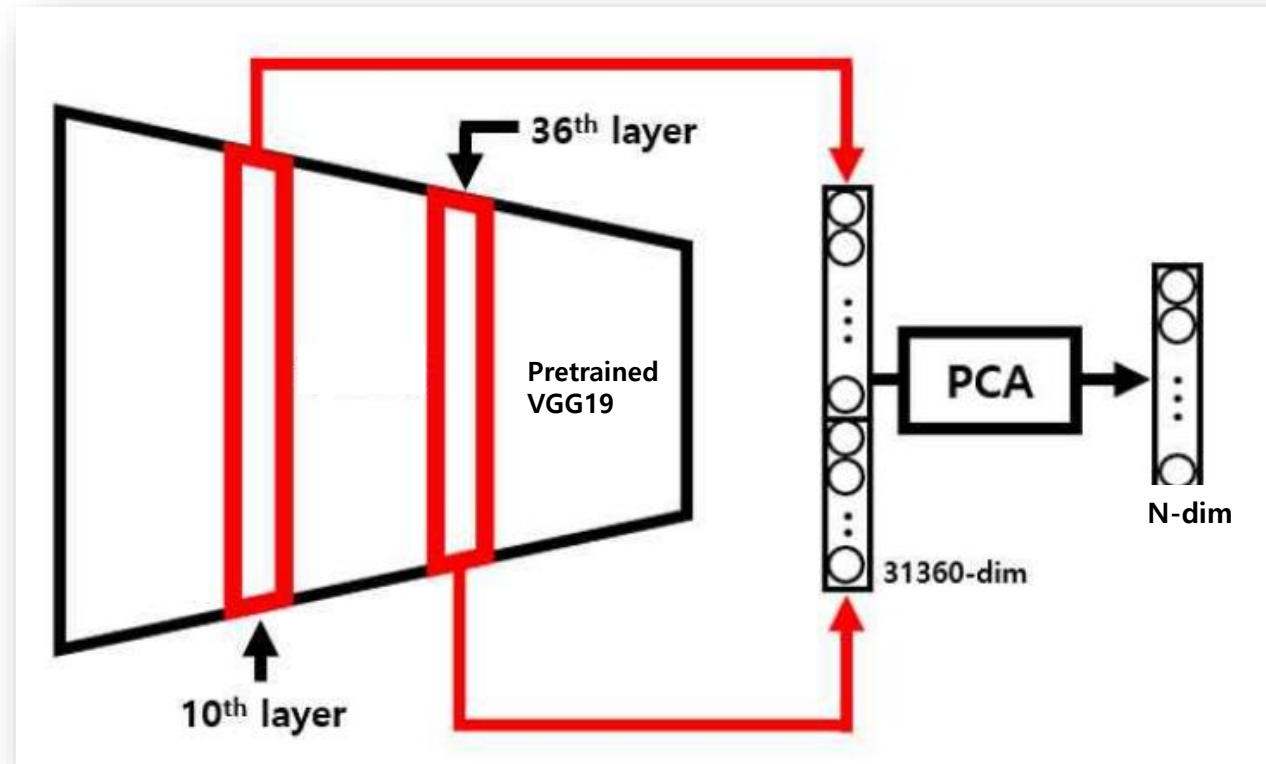
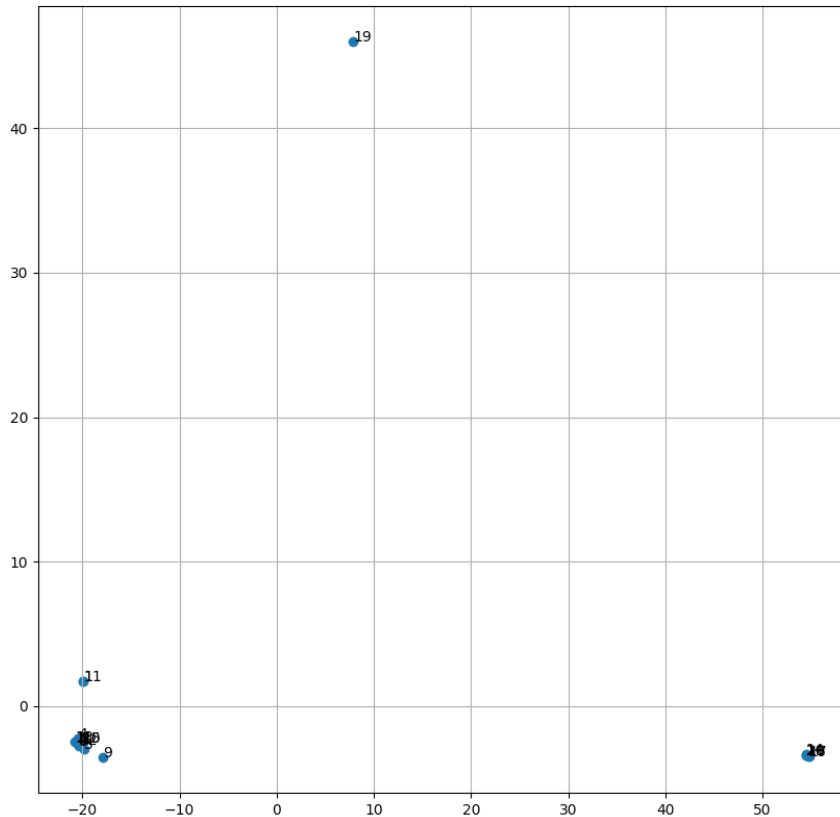
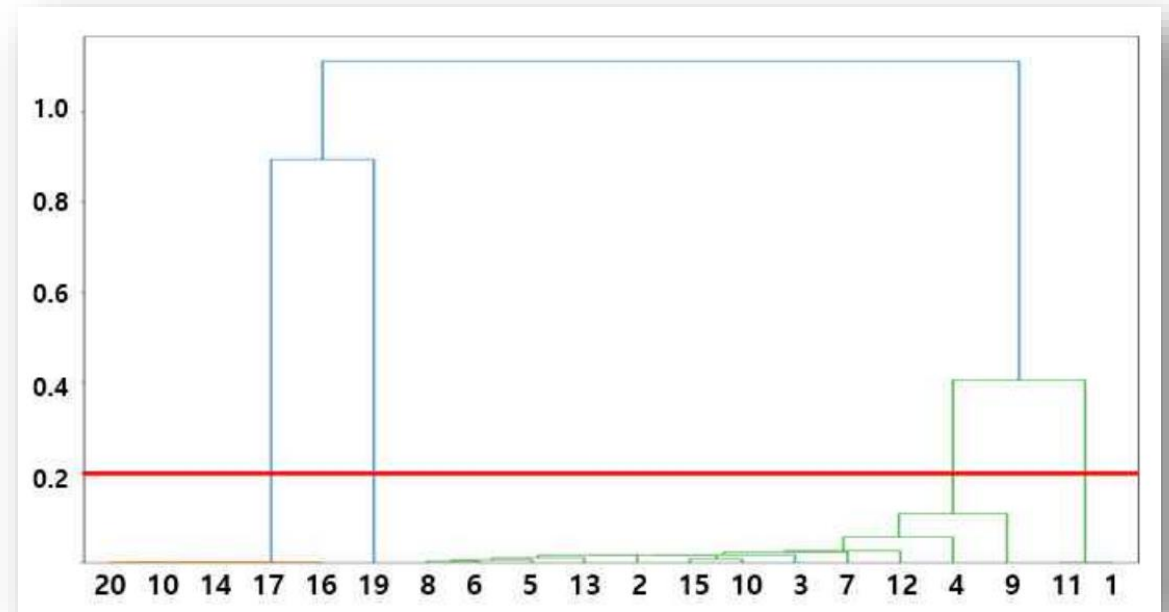


Figure 2. Feature extraction module

## Extracted Features and Clustering Result



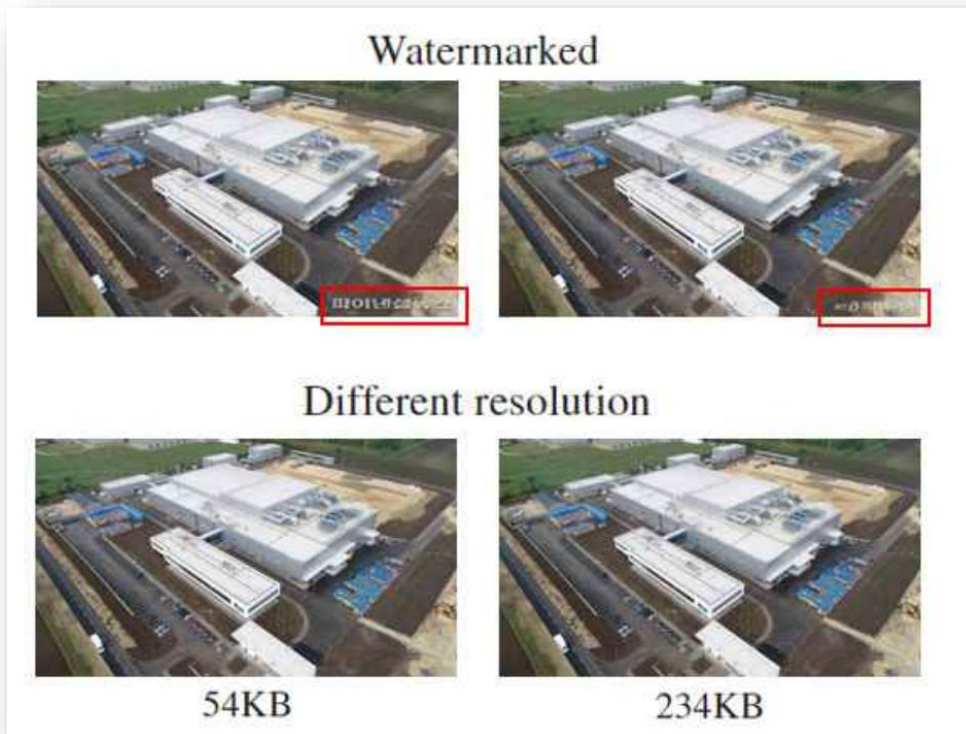
Extracted Features Plotting with PCA



Dendrogram of Hierarchical Clustering

## **IV. Experimental Results**

## Dataset



- 금융당국이 신용대출 규제 강화 카드를 꺼내들었다
- 두산솔루스가 전기차 배터리 핵심소재인 전지박 생산 설비 증설을 승인했다
- 헴데시비르가 미국에서 항바이러스 치료제로 사용할 수 있는 항체신약 헴데시비르를 국내에서 사용할 수 있게 됐다
- 전기차 대중화 시대 성큼 신개념 충전 인프라 개발
- 정부가 공항 입점업체 등에 대한 임대료 감면을 추진한다
- 지난달 서비스 물가 상승률이 0%에 그쳤다
- 한국은행이 기준금리를 사상 처음으로 0%로 인하했다
- 한은은 최근 코로나19 확산세 진정세와 내수 회복으로 인해 국내 경제활동이 다소 위축되는 조짐을 보이고 있다고 진단했다
- 한진택배, 노조 파업에 집하금지 울산 포함 8곳 물류 차질 현실화

9 Queries  
20 Images per query

Figure 3. Example of Duplicated images in the dataset



## Evaluation Criteria

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

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

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

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

$FP$  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

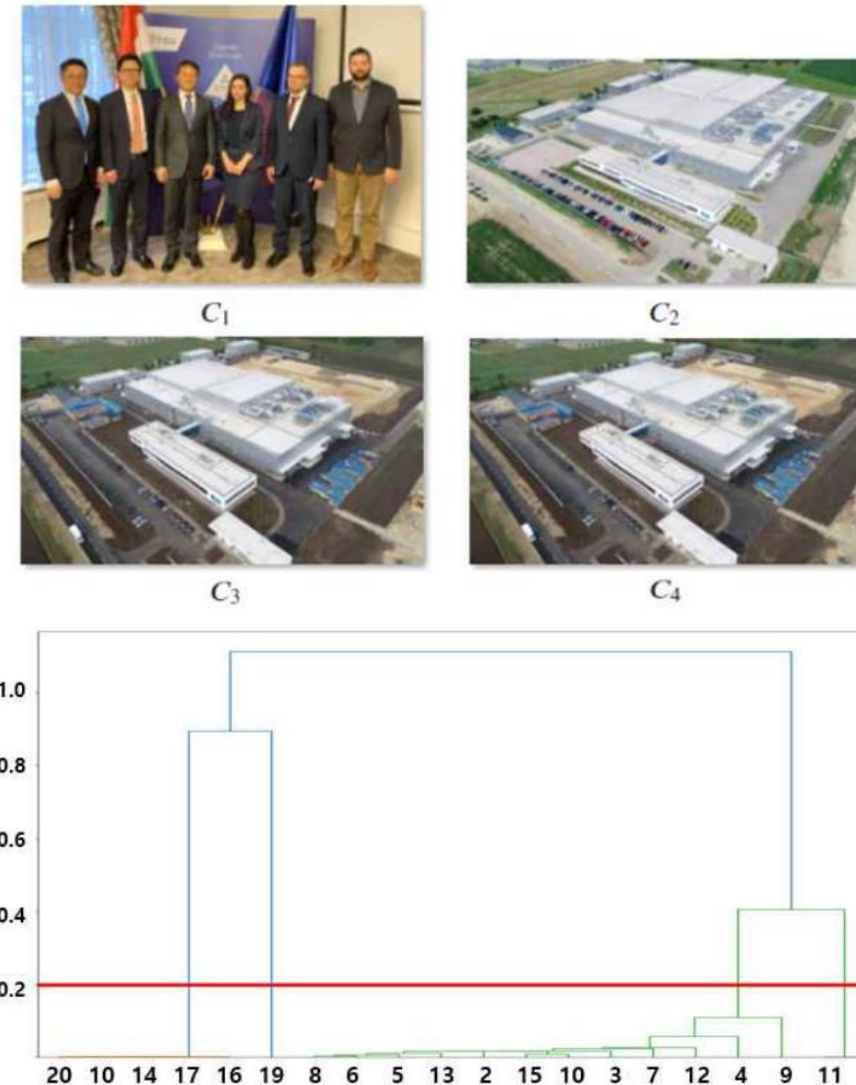
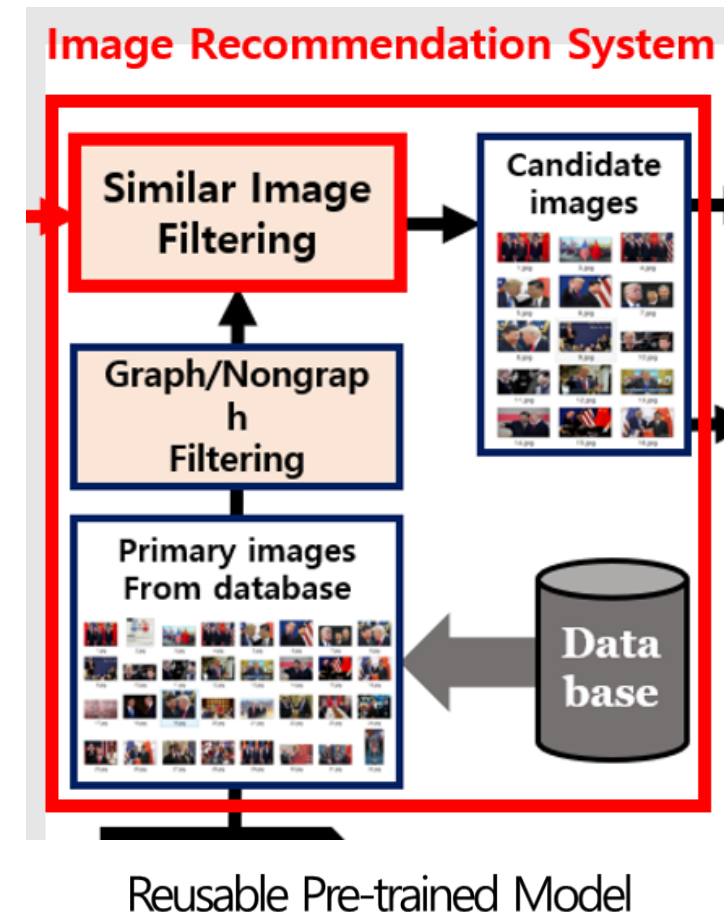
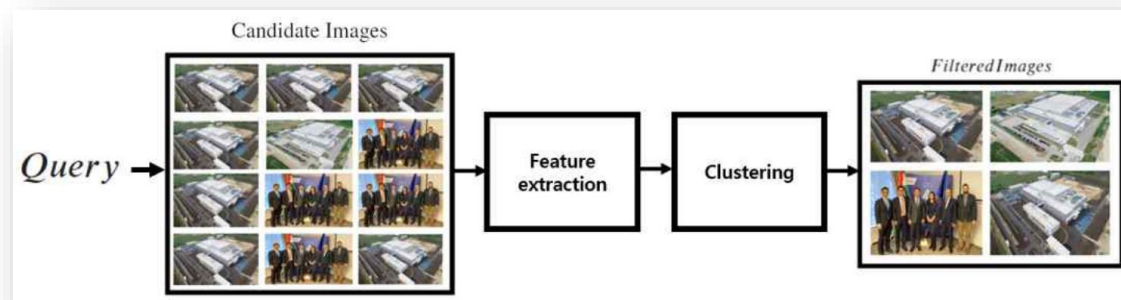


Figure 5. Result of second query



# V. Conclusion

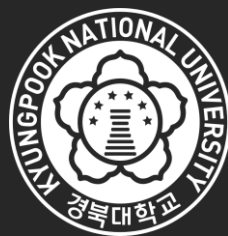
# Conclusion



## **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**