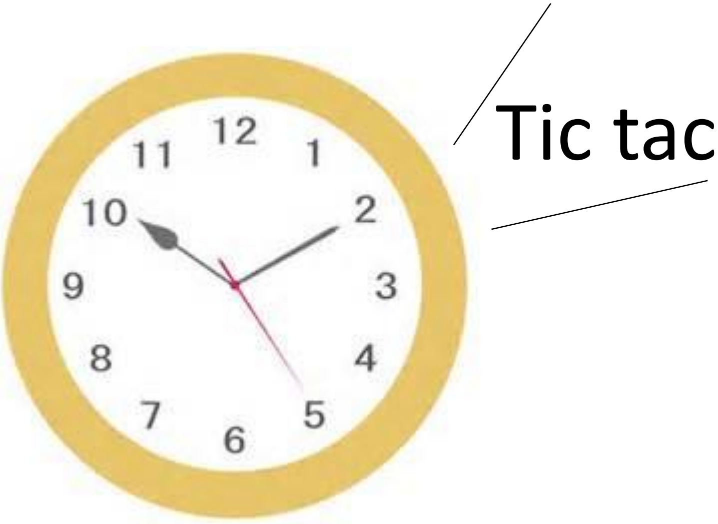
AN ANALYSIS ON VISUAL RECOGNIZABILITY OF ONOMATOPOEIA USING WEB IMAGES AND DCNN FEATURES

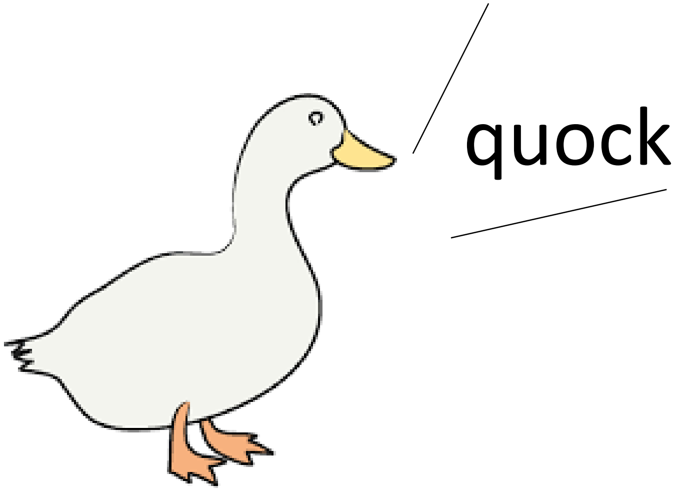
Wataru Shimoda, Yanai Keiji

The Univ. of Electro-Communications Tokyo, Japan

1 INTRODUCTION

- Onomatopoeia
  - Express source of the sound
  - Ex) Tic tac, quock





1 INTRODUCTION

- Onomatopoeia in Japanese
  - Express not only source of sounds
  - Express feeling of visual appearance or touch of objects or materials
  - Many onomatopoeia words

EXAMPLE OF ONOMATOPOEIA IN JAPANESE

fuwa-fuwa
means being very softy
like very soft cotton

zara-zara
means being rough surface
like sandy texture

1 INTRODUCTION

This work:

- Analyze the relation between images and onomatopoeia
- Use a large number of tagged images on the Web
- State-of-the-art visual recognition method
  - Improved Fisher Vector(IFV)
  - Deep Convolutional Neural Network Features (DCNN features)

2 RELATED WORK

- material recognition
  - Flickr Material Database (FMD)
  - Describable Textures Dataset (DTD)
- IFV and DCNN features are effective

FMD image

DTD image

2 RELATED WORK

- Image filtering
  - Amazon Mechanical Turk (AMT)
- AMT has some demerits
  - It costs much
  - To annotate Japanese onomatopoeia is hard for general AMT worker

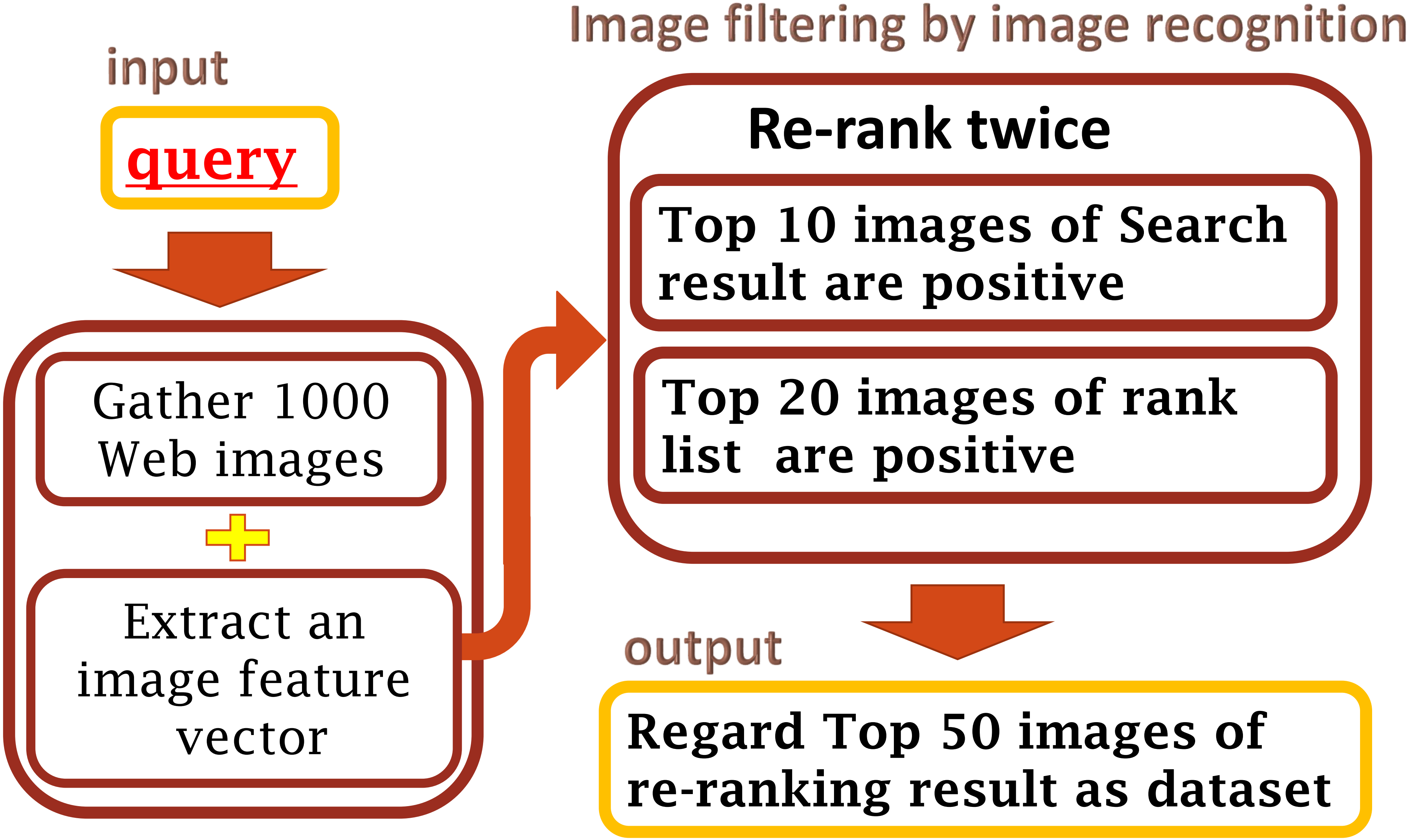
This work:

Constructs an onomatopoeia image dataset based on **automatic** method

3 PROPOSED METHODS

- Construction of onomatopoeia dataset
- Evaluation of gathered onomatopoeia images in terms of recognizability

FLOW OF CONSTRUCTING DATASET



input

query

Gather 1000 Web images

+

Extract an image feature vector

Image filtering by image recognition

Re-rank twice

Top 10 images of Search result are positive

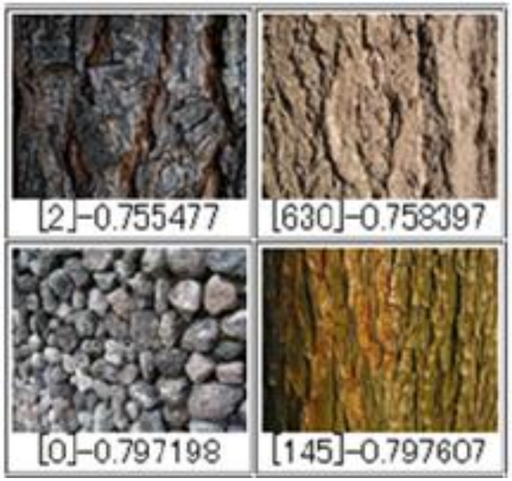
Top 20 images of rank list are positive

output

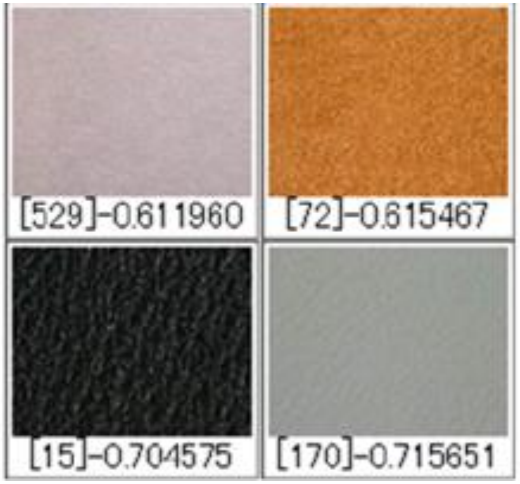
Regard Top 50 images of re-ranking result as dataset

3.1 GATHER WEB IMAGES

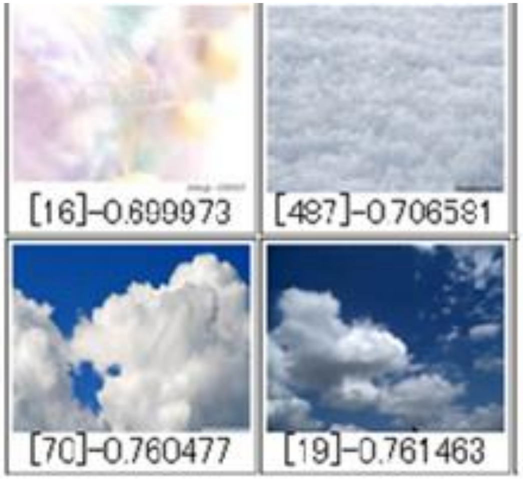
- Bing Image Search API
  - Japanese Onomatopoeia word as query



gotsu-gotus



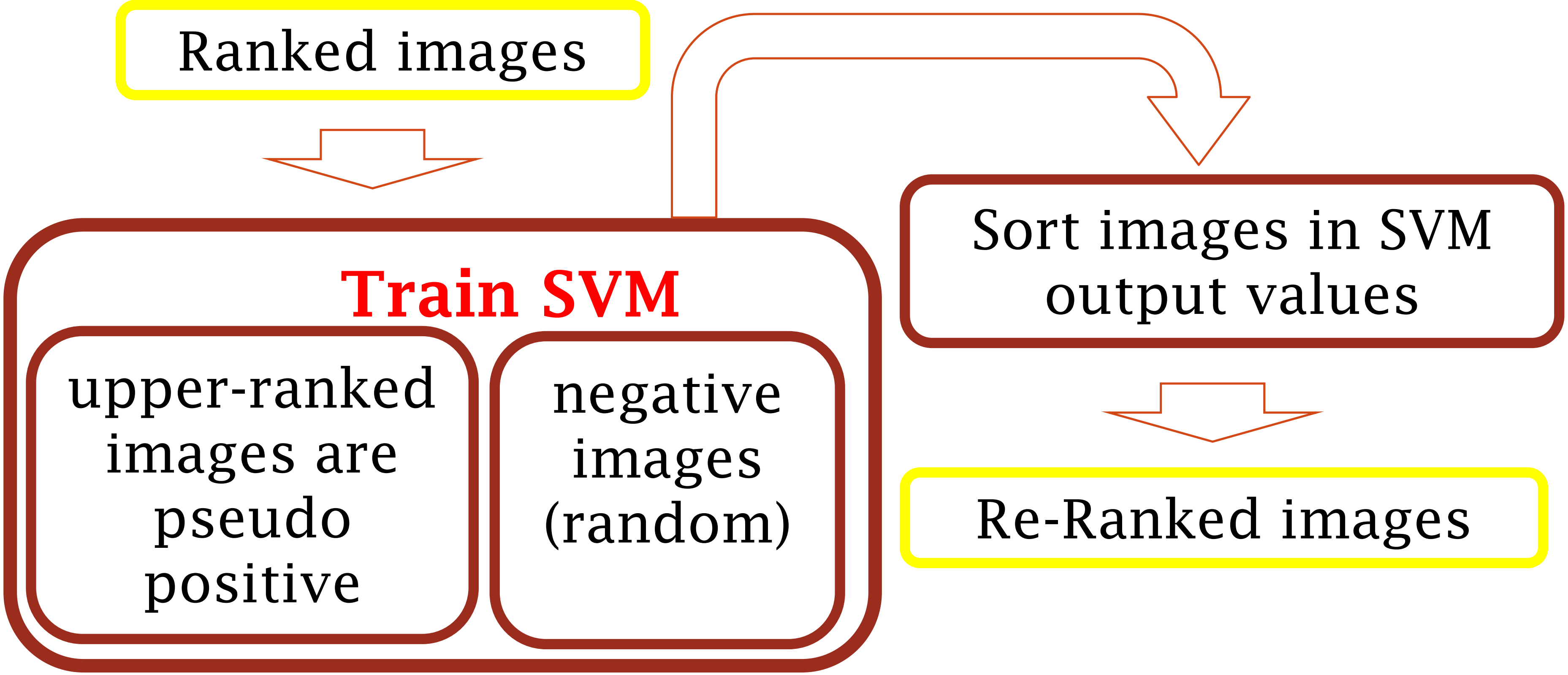
zara-zara



fuwa-fuwa

IMAGE FILTERING

- Re-rank by image recognition



Ranked images

Train SVM

upper-ranked images are pseudo positive

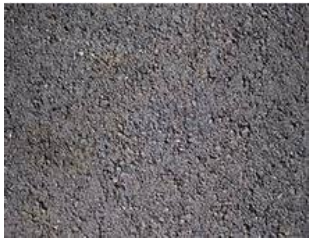
negative images (random)

Sort images in SVM output values

Re-Ranked images

repeat this re-ranking process twice

3.2 RE-RANKING PROCESS DETAIL



expected
zara-zara image

- Gather 1000 image by Bing API

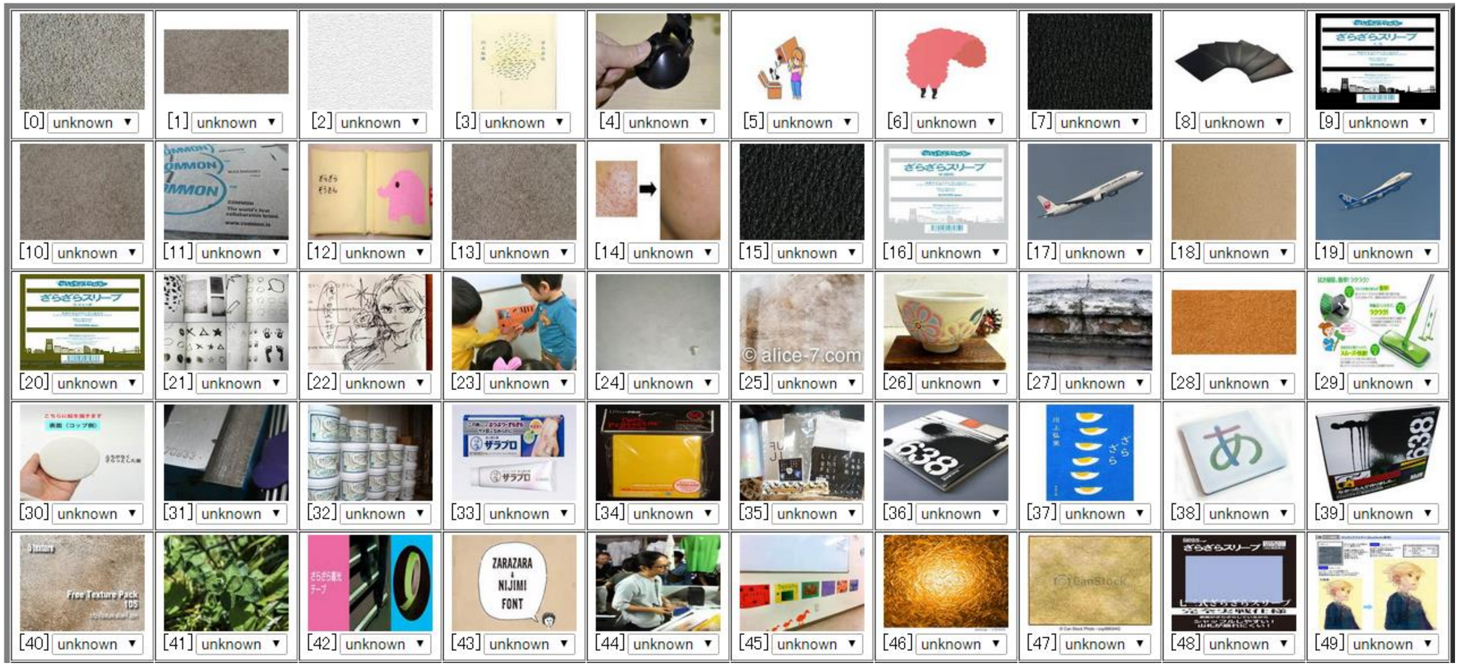
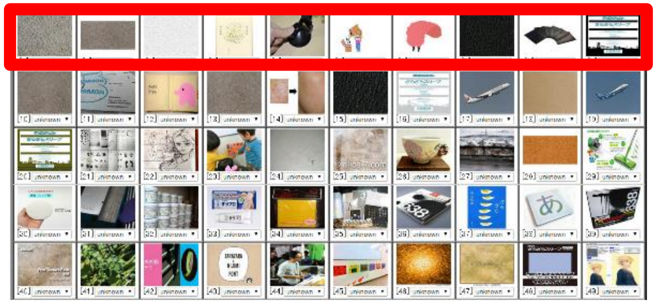


Figure.1
Top 50 image of search result（query: zara-zara）

3.2 RE-RANKING PROCESS DETAIL



- First re-ranking: uses top 10 images of search result as positive images

Figure.1

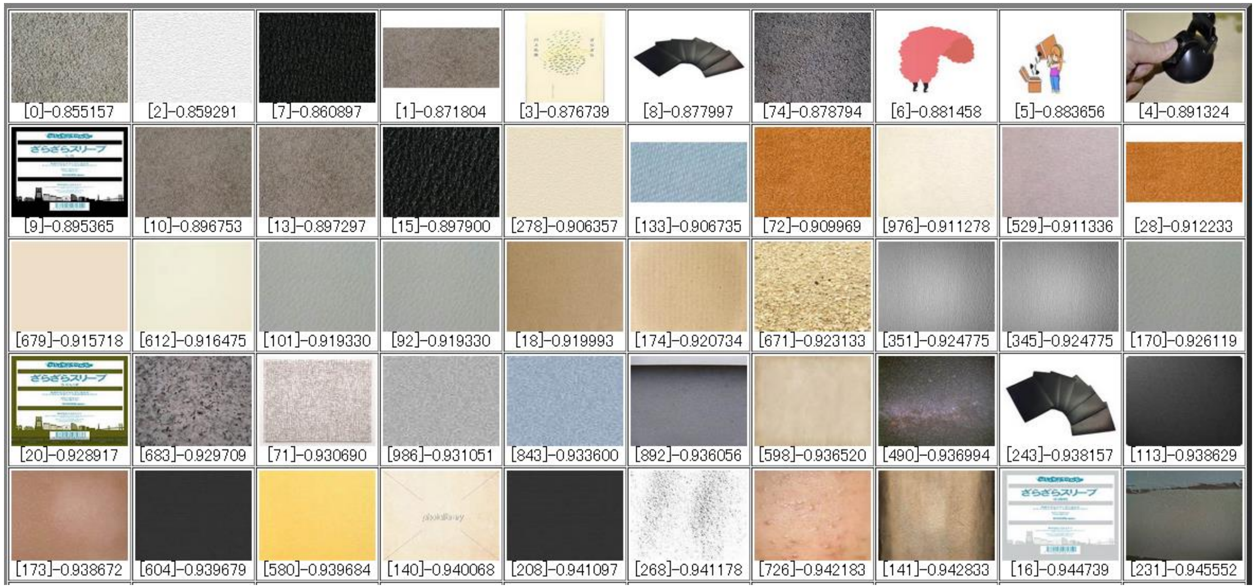


Figure.2
Top 50 image of first re-ranking result（query: zara-zara）

3.2 RE-RANKING PROCESS DETAIL

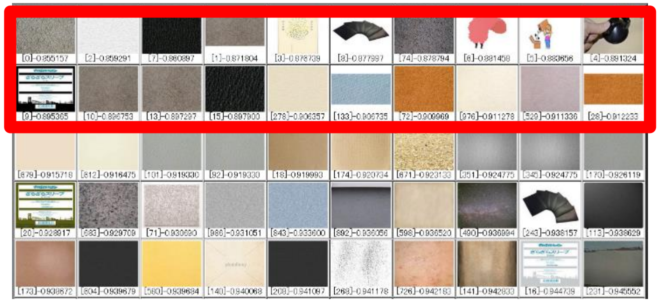


Figure.2

- Second re-ranking: uses top 20 images of first re-ranking result as positive images

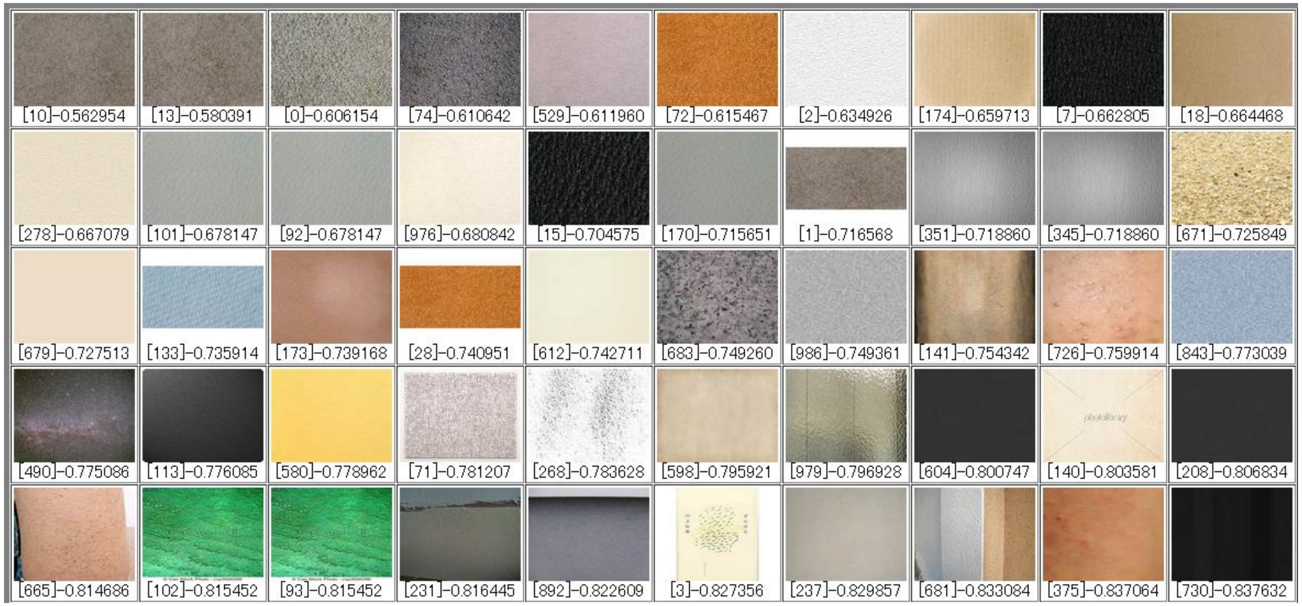


Figure.3
Top 50 image of second re-ranking result （query: zara-zara）

3.3 EVALUATION OF RECOGNIZABILITY OF ONOMATOPOEIA WORDS

- Mix 50 onomatopoeia images and 5000 random noise images
- Discriminate onomatopoeia images from noise images
- Regard that the obtained average precision means the recognizability

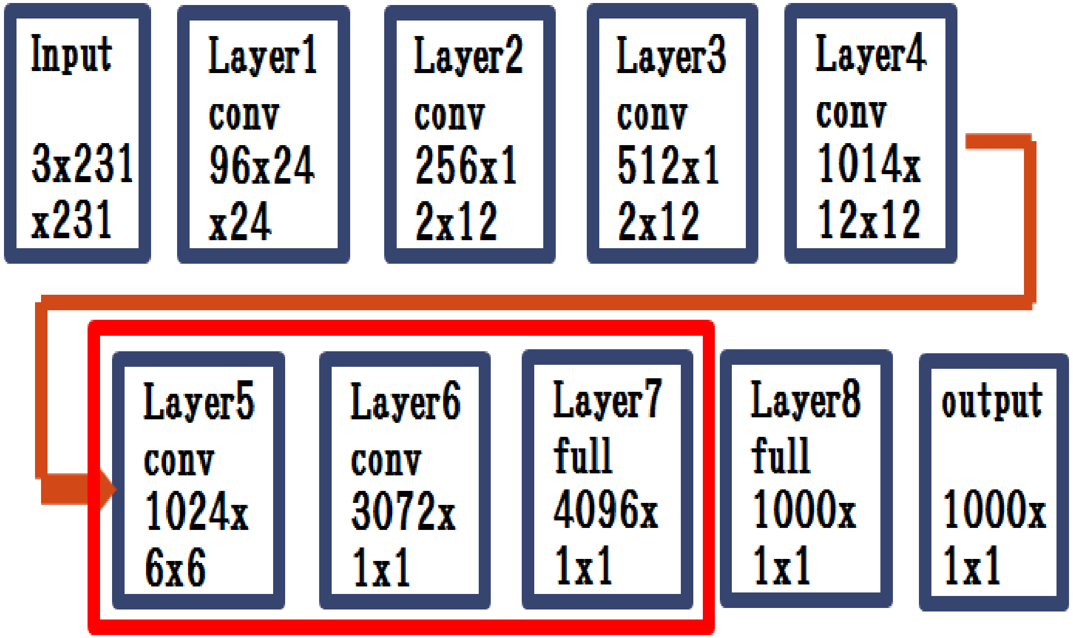
3.4 IMAGE FEATURES

- Image Features
  - Improved fisher vector (IFV)
  - Deep Convolutional Neural Network activation feature (DCNN)

DEEP CONVOLUTIONAL NEURAL NETWORK FEATURES (DCNN FEATURES)

- Overfeat
  - Pre-trained with Image Net 1000 category
  - Use middle layers (layer 5, 6 and 7)
  - L2-normalize

Layer5:
36864 dimension
Layer6:
3072 dimension
Layer7:
4096 dimension



Input 3x231 x231

Layer1 conv 96x24 x24

Layer2 conv 256x1 2x12

Layer3 conv 512x1 2x12

Layer4 conv 1014x 12x12

Layer5 conv 1024x 6x6

Layer6 conv 3072x 1x1

Layer7 full 4096x 1x1

Layer8 full 1000x 1x1

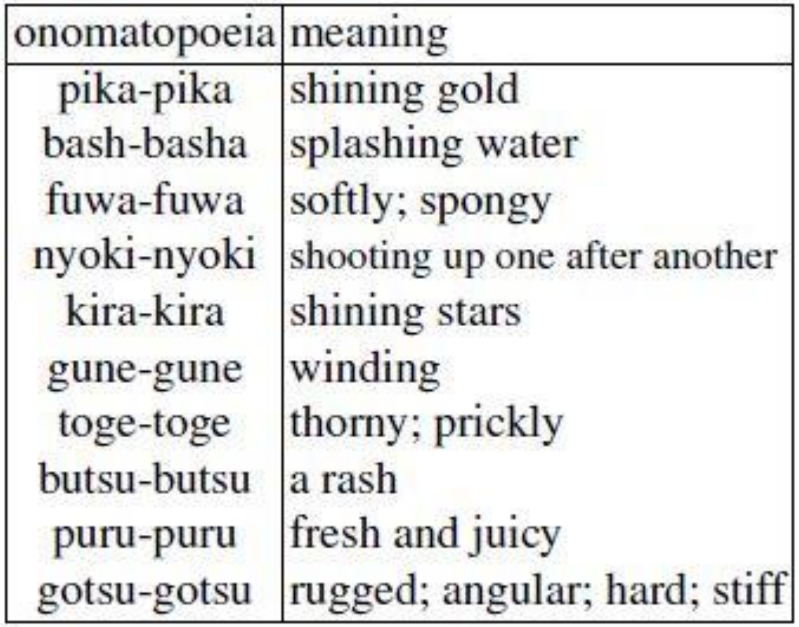
output 1000x 1x1

3.5 CLASSIFICATION

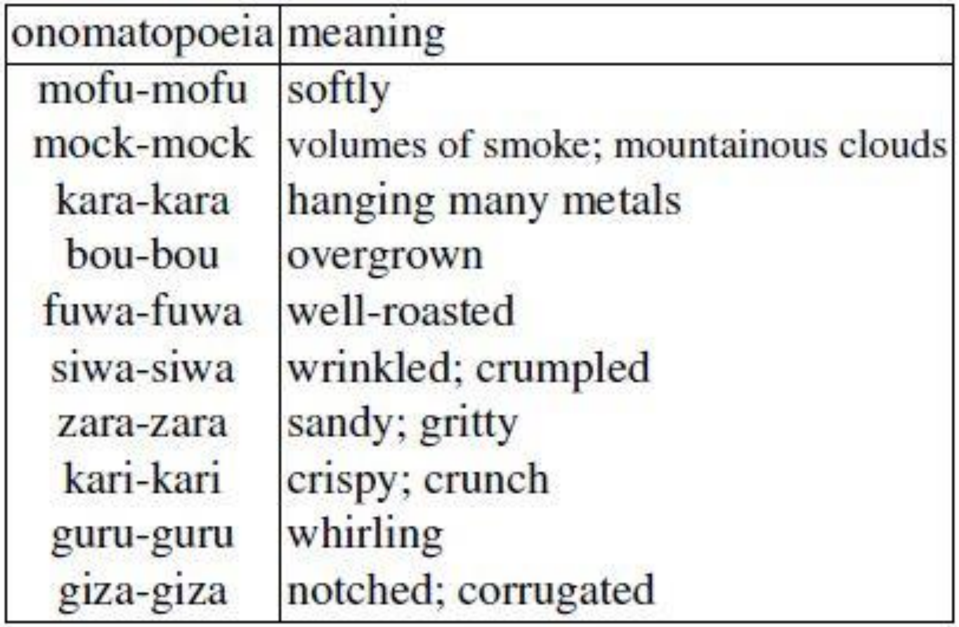
- Support vector machine (SVM)
  - Linear SVM

4 EXPERIMENTS

- Twenty Japanese onomatopoeia words



onomatopoeia	meaning
pika-pika	shining gold
bash-basha	splashing water
fuwa-fuwa	softly; spongy
nyoki-nyoki	shooting up one after another
kira-kira	shining stars
gune-gune	winding
toge-toge	thorny; prickly
butsu-butsu	a rash
puru-puru	fresh and juicy
gotsu-gotsu	rugged; angular; hard; stiff



onomatopoeia	meaning
mofu-mofu	softly
mock-mock	volumes of smoke; mountainous clouds
kara-kara	hanging many metals
bou-bou	overgrown
fuwa-fuwa	well-roasted
siwa-siwa	wrinkled; crumpled
zara-zara	sandy; gritty
kari-kari	crispy; crunch
guru-guru	whirling
giza-giza	notched; corrugated

Zara-zara

Guru-guru

Kari-kari

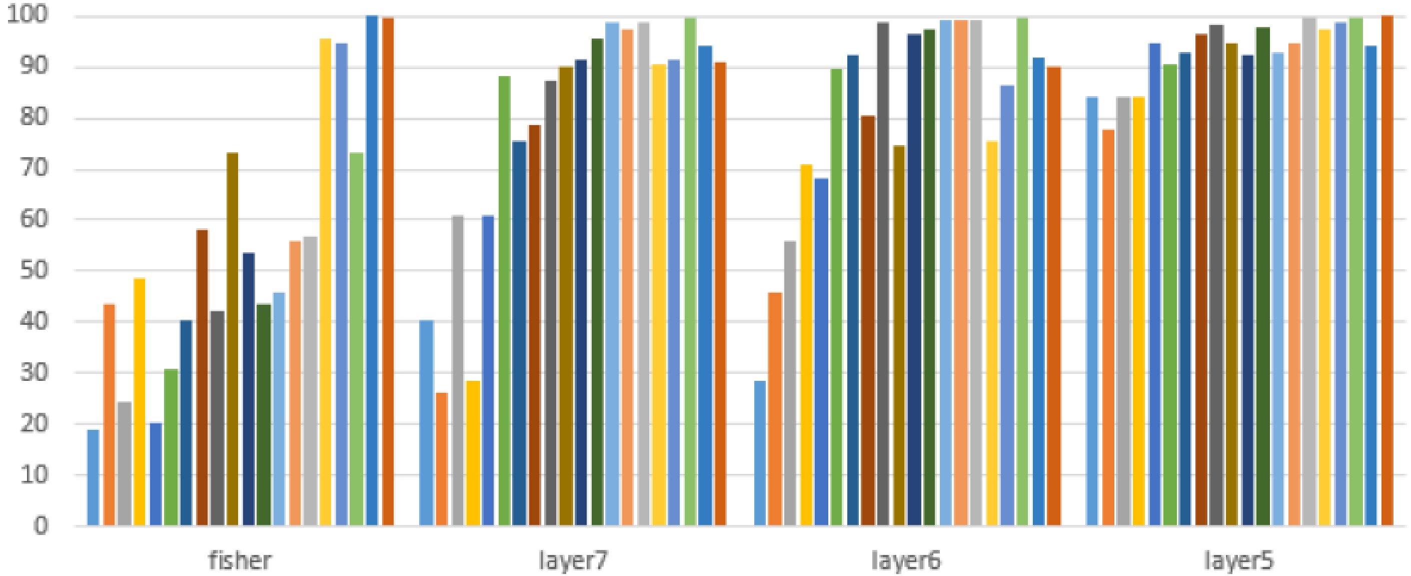
Mock-mock

4.1 EVALUATION OF GATHERED IMAGES

feature Re-ranking	IFV	DCNN		
		Layer7	Layer6	Layer5
Before (search result)	68.6			
After (dataset)	56.0	79.3	82.0	93.2
After-Before (effect(up))	−12.6	+10.7	+13.4	+24.6

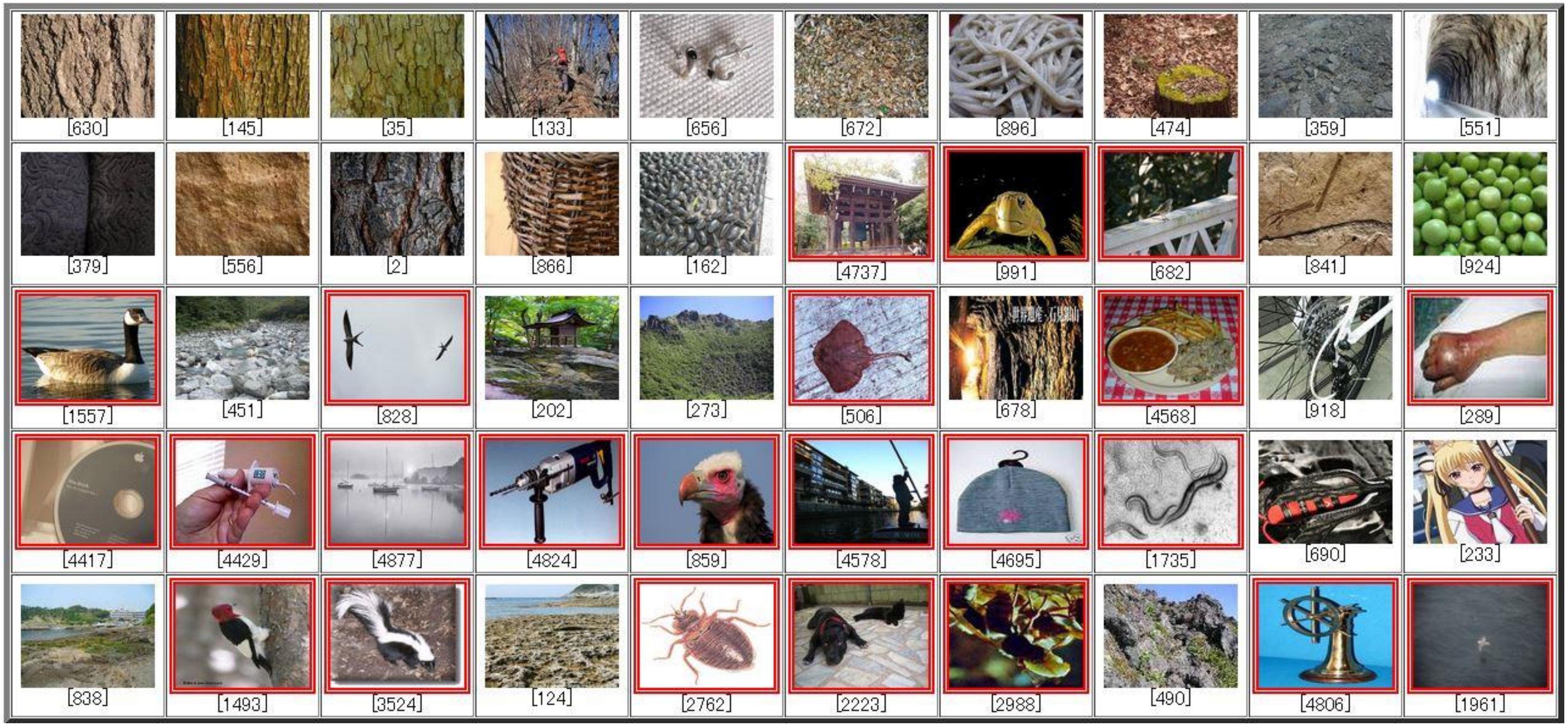
4.2 EVALUATION OF RECOGNIZABILITY

- DCNN features outperformed IFV clearly
- Layer5 result is prominent



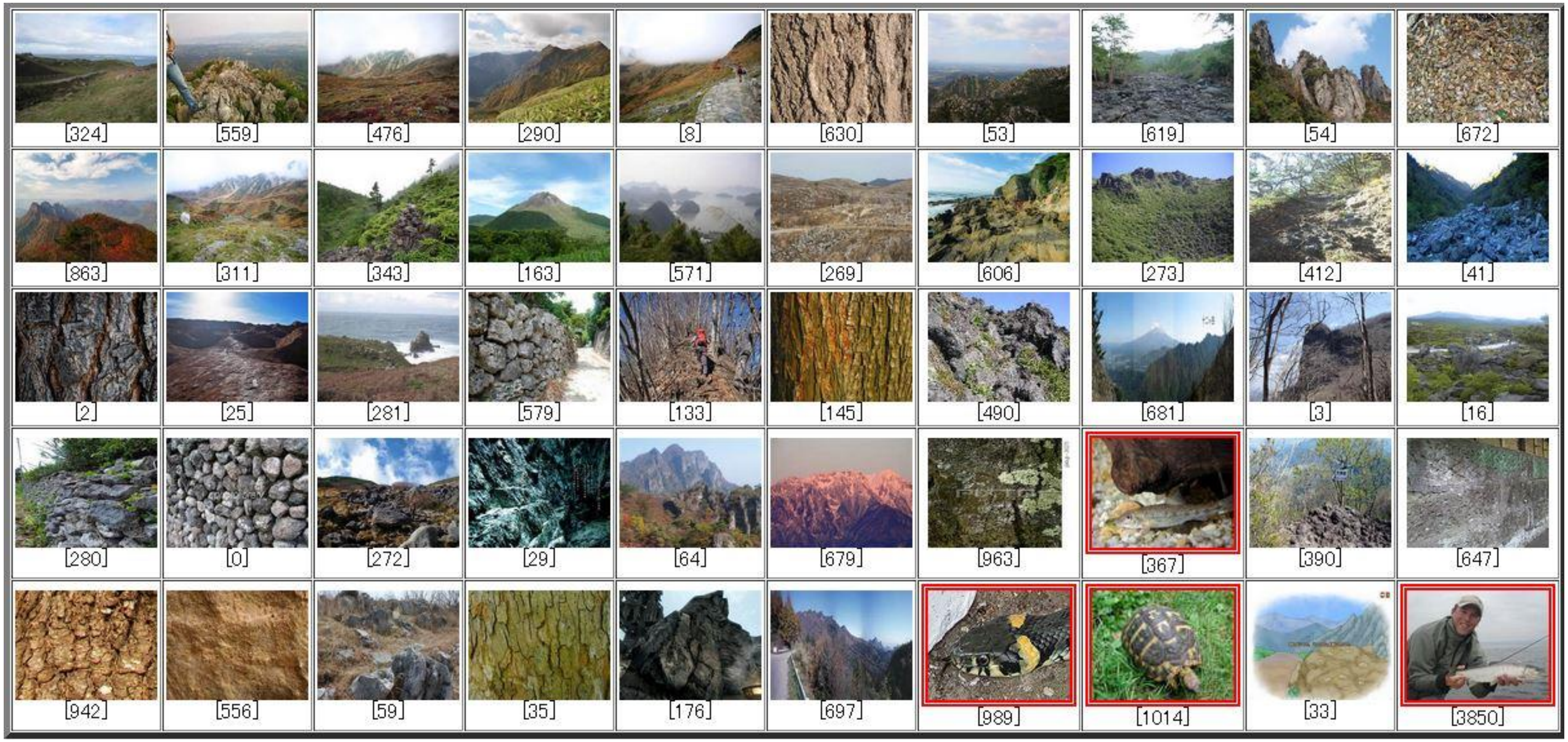
Feature	IFV	Layer7(DCNN)	Layer6(DCNN)	Layer5(DCNN)
Maps(%)	56.0	79.3	82.0	93.2

RECOGNIZABILITY RESULT



IFV （gotsu-gotsu） 73.3%

RECOGNIZABILITY RESULT



DCNN　Layer5　（gotsu-gotsu）　94.5%

5 CONCLUSIONS

- Examined if Japanese onomatopoeia images can be recognized
- DCNN features extracted from the layer 5 achieved 93.2 % maps
- Layer 5 was the most effective feature for onomatopoeia images

END

FUTURE WORK

- Noun + onomatopoeia word
  - Ex) dog + huwa-huwa, dog + shiwa-shiwa
  - onomatopoeia images classification

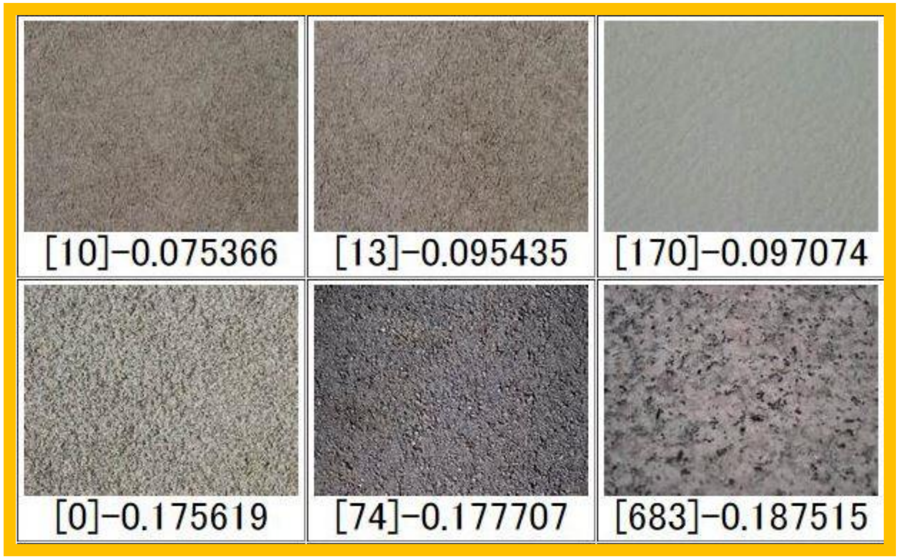
EVALUATE DCNN LAYER PRECISION

- DCNN Layer5 feature result is good
- Not all twenty Onomatopoeia precision is improved

- Improved
  - zara-zara, siwa-siwa → Texture image
- Not improved
  - jara-jara, mohu-mohu → object image

IMPROVED BY LAYER5 FEATURE

- Texture image



zara-zara

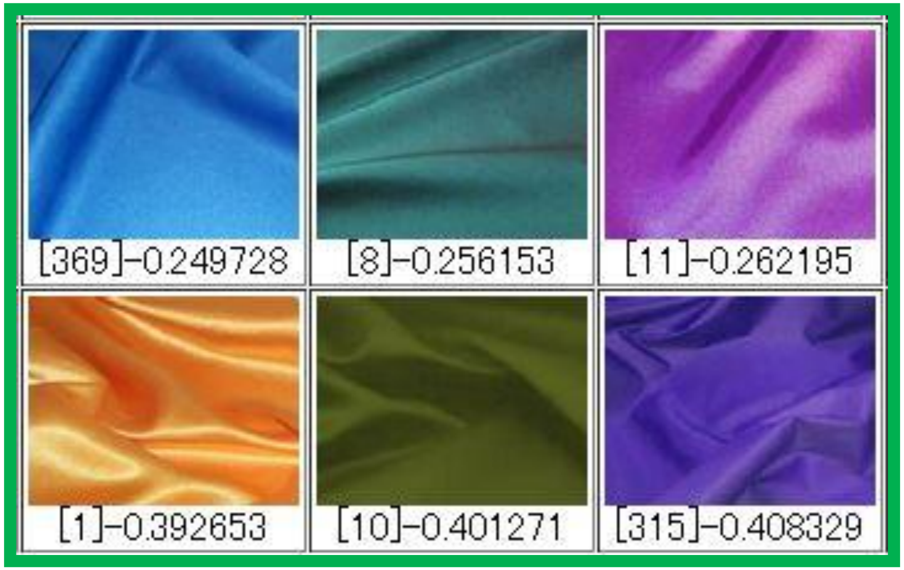
Layer6: 86.4%

Layer5: 98.7% **+12.3%**

shiwa-shiwa

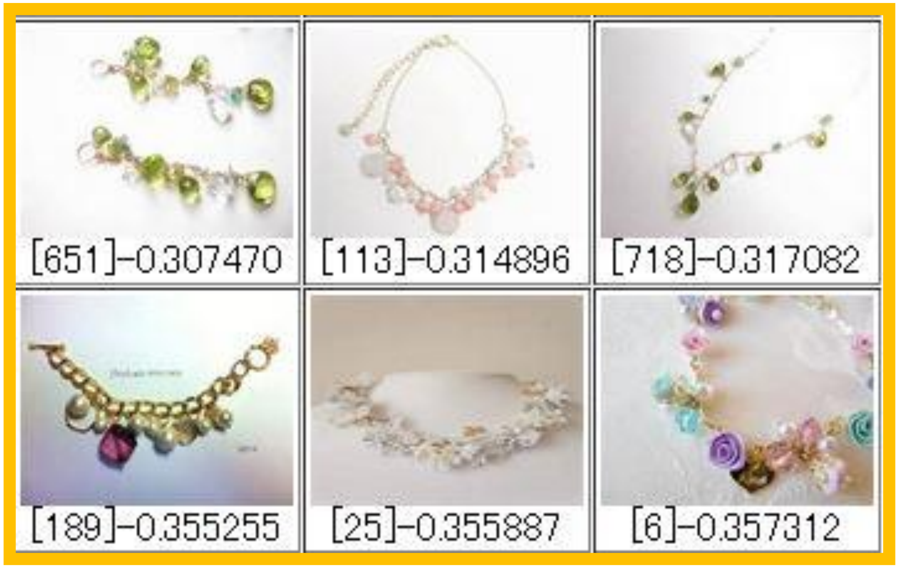
Layer6: 75.5%

Layer5: 97.6% **+22.1%**



NOT IMPROVED BY LAYER5 FEATURE

- Object image



jara-jara

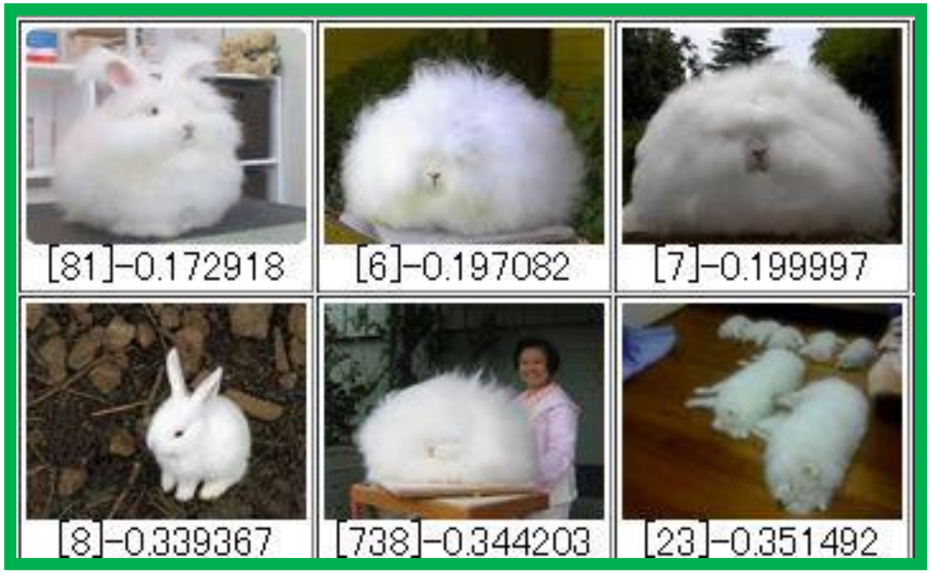
Layer6: 99.4%

Layer5: 92.7% **-6.7%**

mofu-mofu

Layer6: 96.4%

Layer5: 92.4% **-6.0%**



FEATURE MAPS

- Layer6 and Layer 7 precision is improved by feature maps

Feature	DCNN		
	Feature maps		Layer5
	Layer7	Layer6	
Maps(%)	**91.3**	**95.3**	**93.2**

NEGATIVE IMAGE

- Image net
  - 10,000 category
  - We gather an one image each category

- We use the same feature in the two steps re-ranking and evaluating
- IFV can fail to construct the dataset.
- IFV precision may be reduced excessively by the method

SVM

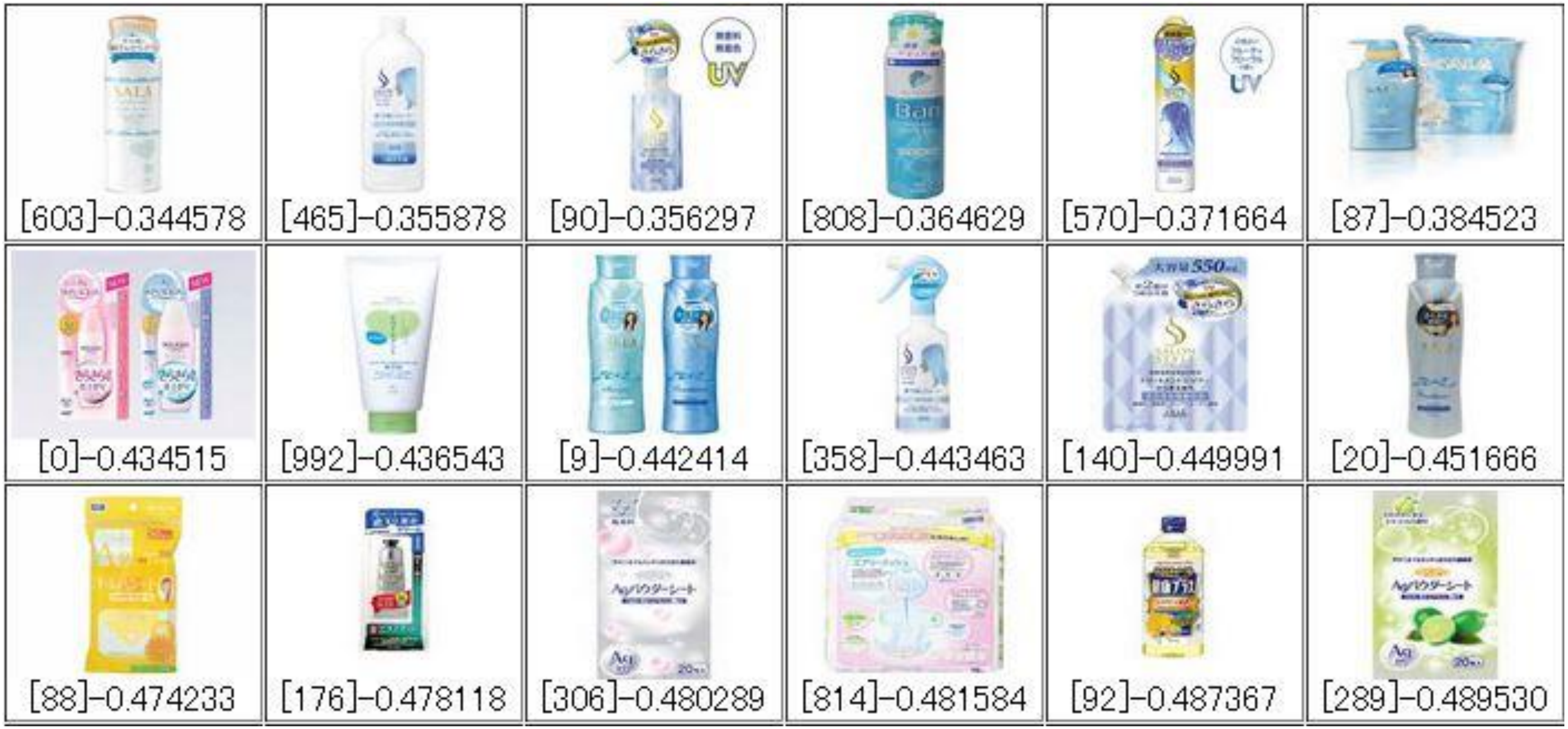
- SVM train with 50 positive images + 1000 negative images

- Use another 5000 negative images to evaluate recognizability

FAILED CASE

- Sara-sara

We expected such a sara-sara object


[603]-0.344578
[465]-0.355878
[90]-0.356297
[808]-0.364629
[570]-0.371664
[87]-0.384523
[0]-0.434515
[992]-0.436543
[9]-0.442414
[358]-0.443463
[140]-0.449991
[20]-0.451666
[88]-0.474233
[176]-0.478118
[306]-0.480289
[814]-0.481584
[92]-0.487367
[289]-0.489530