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

zara-zara
means being rough surface like sandy texture

fuwa-fuwa
means being very softly like very soft cotton
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
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

![Images of various textures and patterns with associated Japanese sound effects: gotsu-gotus, zara-zara, fuwa-fuwa]
IMAGE FILTERING

- Re-rank by image recognition

- 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

- 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.2
Top 50 image of first re-ranking result (query: zara-zara)
3.2 RE-RANKING PROCESS DETAIL

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

Figure.2
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)
Overfeat
- Pre-trained with Image Net 1000 category
- Use middle layers (layer 5, 6 and 7)
- L2-normalize

Layer 5:
36864 dimension

Layer 6:
3072 dimension

Layer 7:
4096 dimension
3.5 CLASSIFICATION

- Support vector machine (SVM)
  - Linear SVM
4 EXPERIMENTS

- Twenty Japanese onomatopoeia words

<table>
<thead>
<tr>
<th>onomatopoeia</th>
<th>meaning</th>
<th>onomatopoeia</th>
<th>meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>pika-pika</td>
<td>shining gold</td>
<td>mofu-mofu</td>
<td>softly</td>
</tr>
<tr>
<td>bash-basha</td>
<td>splashing water</td>
<td>mock-mock</td>
<td>volumes of smoke; mountainous clouds</td>
</tr>
<tr>
<td>fuwa-fuwa</td>
<td>softly; spongy</td>
<td>kara-kara</td>
<td>hanging many metals</td>
</tr>
<tr>
<td>nyoki-nyoki</td>
<td>shooting up one after another</td>
<td>bou-bou</td>
<td>overgrown</td>
</tr>
<tr>
<td>kira-kira</td>
<td>shining stars</td>
<td>fuwa-fuwa</td>
<td>well-roasted</td>
</tr>
<tr>
<td>gune-gune</td>
<td>winding</td>
<td>siwa-siwa</td>
<td>wrinkled; crumpled</td>
</tr>
<tr>
<td>toge-toge</td>
<td>thorny; prickly</td>
<td>zara-zara</td>
<td>sandy; gritty</td>
</tr>
<tr>
<td>butsu-butsu</td>
<td>a rash</td>
<td>kari-kari</td>
<td>crispy; crunch</td>
</tr>
<tr>
<td>puru-puru</td>
<td>fresh and juicy</td>
<td>guru-guru</td>
<td>whirling</td>
</tr>
<tr>
<td>gotsu-gotsu</td>
<td>rugged; angular; hard; stiff</td>
<td>giza-giza</td>
<td>notched; corrugated</td>
</tr>
</tbody>
</table>

Zara-zara  Guru-guru  Kari-kari  Mock-mock
## 4.1 Evaluation of Gathered Images

<table>
<thead>
<tr>
<th>Re-ranking</th>
<th>Feature</th>
<th>IFV</th>
<th>DCNN</th>
<th>Layer7</th>
<th>Layer6</th>
<th>Layer5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before (search result)</td>
<td></td>
<td>68.6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>After (dataset)</td>
<td></td>
<td>56.0</td>
<td>79.3</td>
<td>82.0</td>
<td>93.2</td>
<td></td>
</tr>
<tr>
<td>After-Before (effect(up))</td>
<td>-12.6</td>
<td>+10.7</td>
<td>+13.4</td>
<td>+24.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4.2 EVALUATION OF RECOGNIZABILITY

- DCNN features outperformed IFV clearly
- Layer5 result is prominent

<table>
<thead>
<tr>
<th>Feature</th>
<th>IFV</th>
<th>Layer7(^{\text{DCNN}})</th>
<th>Layer6(^{\text{DCNN}})</th>
<th>Layer5(^{\text{DCNN}})</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maps(%)</td>
<td>56.0</td>
<td>79.3</td>
<td>82.0</td>
<td>93.2</td>
</tr>
</tbody>
</table>
RECOGNIZABILITY RESULT

IFV (gotsu–gotsu) 73.3%
RECOGNIZABILITY RESULT

<table>
<thead>
<tr>
<th>324</th>
<th>556</th>
<th>476</th>
<th>290</th>
<th>8</th>
<th>630</th>
<th>53</th>
<th>619</th>
<th>54</th>
<th>672</th>
</tr>
</thead>
<tbody>
<tr>
<td>863</td>
<td>511</td>
<td>343</td>
<td>163</td>
<td>571</td>
<td>269</td>
<td>605</td>
<td>273</td>
<td>412</td>
<td>41</td>
</tr>
<tr>
<td>2</td>
<td>25</td>
<td>281</td>
<td>579</td>
<td>133</td>
<td>145</td>
<td>480</td>
<td>681</td>
<td>5</td>
<td>16</td>
</tr>
<tr>
<td>280</td>
<td>0</td>
<td>272</td>
<td>29</td>
<td>64</td>
<td>679</td>
<td>563</td>
<td>367</td>
<td>350</td>
<td>647</td>
</tr>
<tr>
<td>942</td>
<td>556</td>
<td>59</td>
<td>35</td>
<td>176</td>
<td>87</td>
<td>88</td>
<td>1014</td>
<td>88</td>
<td>3850</td>
</tr>
</tbody>
</table>

**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

- Not improved
  - jara-jara, mohu-mohu

Texture image

object image
IMPROVED BY LAYER 5 FEATURE

- Texture image

**shiwa-shiwa**
Layer6: 75.5%
Layer5: 97.6% \(+22.1\%\)

**zara-zara**
Layer6: 86.4%
Layer5: 98.7% \(+12.3\%\)
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**

- Layer 6 and Layer 7 precision is improved by feature maps

<table>
<thead>
<tr>
<th>Feature</th>
<th>DCNN</th>
<th>Layer5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature maps</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Layer7</td>
<td>Layer6</td>
<td></td>
</tr>
<tr>
<td>Maps(%)</td>
<td>91.3</td>
<td>95.3</td>
</tr>
</tbody>
</table>
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

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>![Image 37]</td>
<td>![Image 38]</td>
<td>![Image 39]</td>
<td>![Image 40]</td>
<td>![Image 41]</td>
<td>![Image 42]</td>
</tr>
</tbody>
</table>

- [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