Waseda at TRECVID 2016: Fully-automatic Ad-hoc Video Search
Kotaro Kikuchi, Kazuya Ueki, Susumu Saito, Tetsunori Kobayashi
Waseda University

1. Fully-automatic Ad-hoc Video Search

Query
“501 Find shots of a person playing guitar outdoors”

Task:
Given the query, return ranked shots based on the likelihood of containing the target query

- No additional human input
- No additional annotation
- Using available dataset or general search results on the Web

2. System Description

Pipeline processing:
1. Automatically select several concepts based on the word similarity
2. Calculate a score for each concept using visual features
3. Combine the semantic concepts to get the final scores

[ Step 2: Score calculation ]

- Extract visual features from the output layer of pre-trained convolutional neural networks (CNNs)
- Normalize visual features over all the test dataset to use them as the scores

[ Step 1: Concept selection ]

- Lemmatize each word which is in the given query
- Convert lemmas and concept names into word vectors by Word2Vec
- Calculate cosine similarities between each lemma and concept name
- Use concept for next step if its similarity is larger than the threshold

[ Step 3: Score integration ]

Final score is simply calculated by multiplying concept scores

\[ \prod_{i=1}^{N} s_i \]

N: # of words
s_i: shot score for each concept

3. Results

**Good case**

Query
“501 Find shots of a person playing guitar outdoors”

Ranked shots

**Bad case**

Query
“505 Find shots of a person holding a poster on the street at daytime”

Ranked shots

**Except “Hold”**

4. Conclusion

- Our system achieved to retrieve videos fully-automatically by the query phrase.
- The accuracy depends on the degree of mismatch concepts which affect results badly.
- In the future, we will integrate the human inexplicit knowledge into our system.

- Tend to select mismatch concept if the word is transitive verb
- Mismatch concepts affect the result badly