Waseda at TRECVID 2016: Fully-automatic Ad-hoc Video Search

Kotaro Kikuchi, Kazuya Ueki, Susumu Saito, Tetsunori Kobayashi

Waseda University

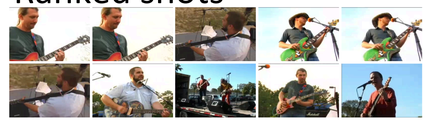
1. Fully-automatic Ad-hoc Video Search

Query

"501 Find shots of a person playing guitar outdoors"

Retrieval system

Ranked shots



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 1

Query

"515 Find shots of a person jumping"

Word	Concept name	Similarity
person	Person	1.000
	Single_Person	0.795
	Young_Person	0.779
jump	Long_Jump	0.835
	High_Jump	0.831

[Step 1: Concept selection]

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

[Step 2: Score calculation]

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

Step 2

Word	Concept name	Shot Score A	Shot Score B
person	Person	0.748	0.753
	Single_Person	0.093	0.100
	Young_Person	0.143	0.127
jump	Long_Jump	0.627	0.318
	High_Jump	0.278	0.112

Step 3

Final score: 0.281 | 0.157

[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

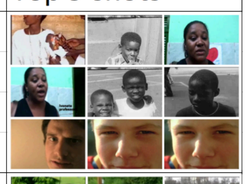
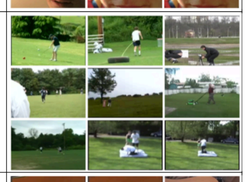
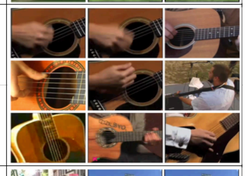
3. Results

Good case

Query

"501 Find shots of a person playing guitar outdoors"

Ranked shots

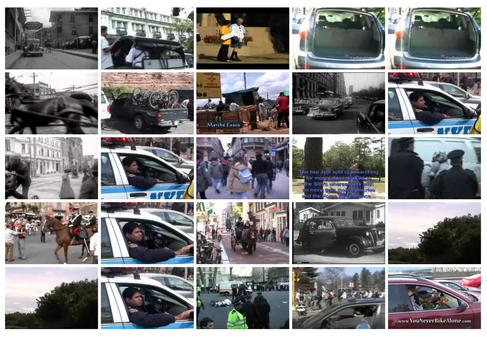
Words	Concept name	Top 9 shots
person	Person	
	Single_Person	
	Young_Person	
	Female_Person	
play	Match_Play	
guitar	Guitar	
	Acoustic_Guitar	
outdoors	Outdoor	

Bad case

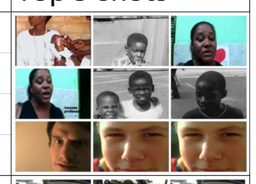
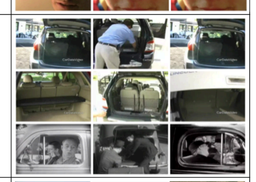
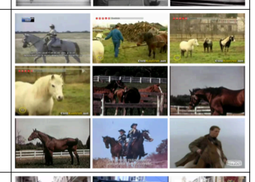
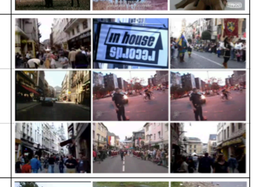
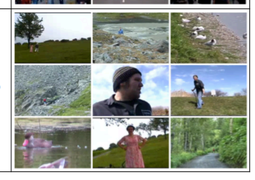
Query

"505 Find shots of a person holding a poster on the street at daytime"

Ranked shots



Except "Hold"

Words	Concept name	Top 9 shots
person	Person	
	Single_Person	
	Young_Person	
	Female_Person	
hold	Hold	
poster	Poster	
street	Street	
	Street_Sign	
	Street_Cloths	
daytime	Daytime_Outdoor	

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

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.