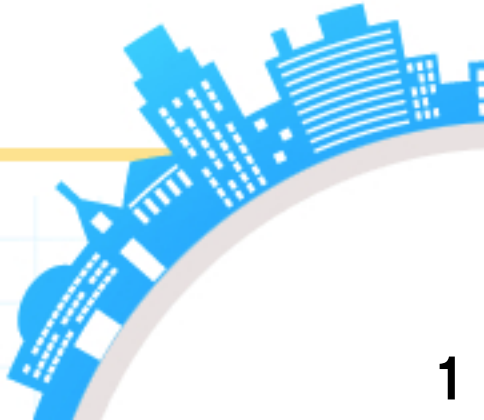
Waseda at TRECVID 2016

Ad-hoc Video Search (AVS)

Kazuya UEKI　　Kotaro KIKUCHI
Susumu SAITO　Tetsunori KOBAYASHI

Waseda University



Outline

1. Introduction
2. System description
3. Submission
4. Results
5. Summary and future works

1. Introduction

1. Introduction

Ad-hoc Video Search (AVS)　Manually assisted runs

Ad-hoc query:

"Find shots of any type of fountains outdoors"

Manually select some keywords.

fountain　outdoor

System takes search keywords and produces results.

Search results

2. System description

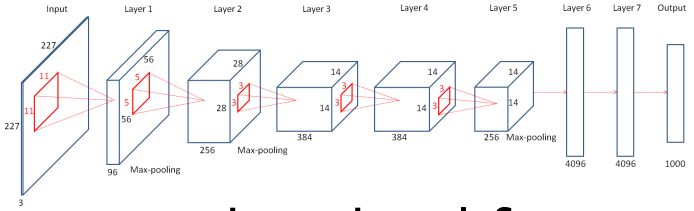
2. System description

Our method consists of three steps:

[Step. 1]

Manually select several search keywords based on the given query phrase.

[Step. 2]



Calculate a score for each concept using visual features.

[Step. 3]

Combine the semantic concepts to get the final scores.

2. System description

[Step. 1]

Manually select several search keywords based on the given query phrase.

We explicitly distinguished *and* from *or*.

Example 1

"any type of fountains outdoors"

⇒ "fountain" and "outdoor"

Example 2

"one or more people walking or bicycling on a bridge during daytime"

⇒ "people" and ("walking" or "bicycling") and "bridge" and "daytime"

2. System description

[Step. 2]

Calculate a score for each concept using visual features.

We extracted visual features from pre-trained convolutional neural networks (CNNs)

Pre-trained models used in our runs

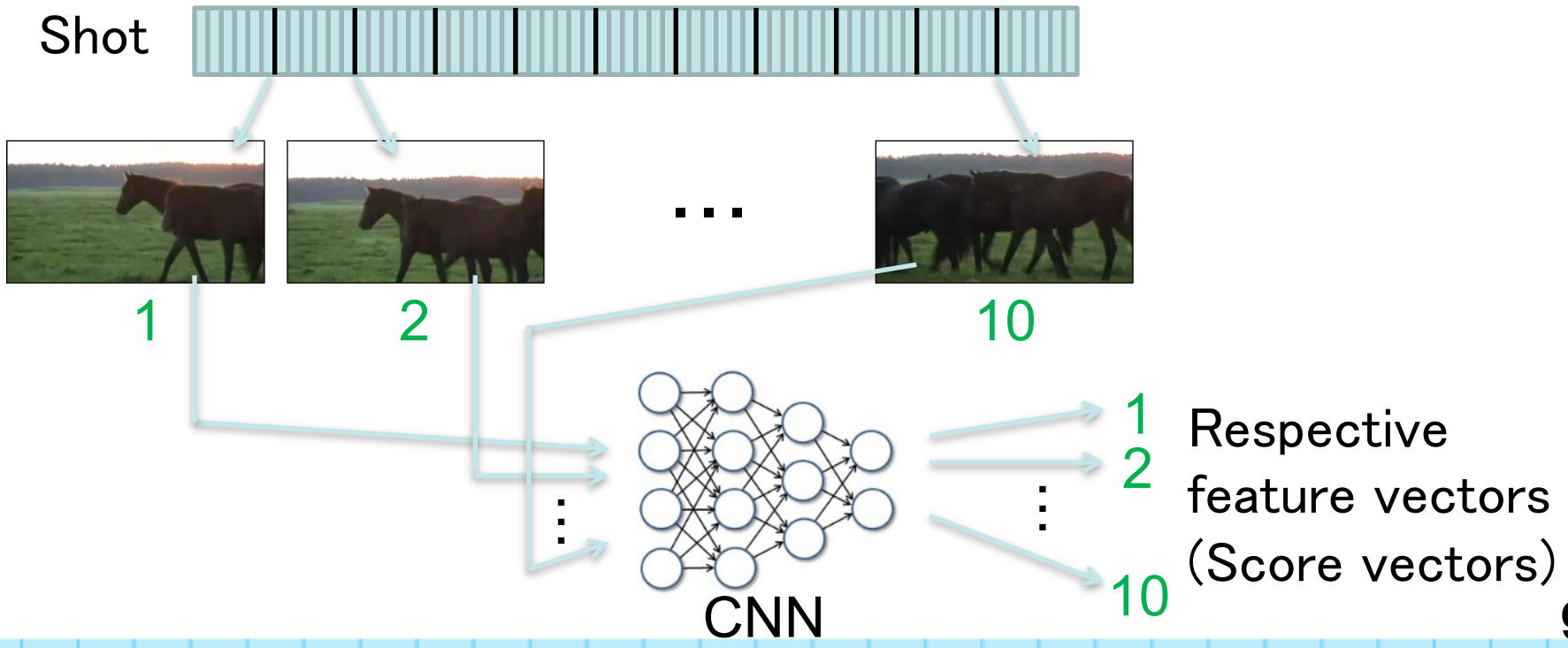
Model name	Database	Number of concepts	Concept type(s)
TRECVID346	TRECVID (ImageNet)	346	Object, Scene, Action
PLACES205	Places	205	Scene
PLACES365	Places	365	Scene
HYBRID1183	Places, ImageNet	1,183	Object, Scene
IMAGENET1000	ImageNet	1,000	Object
IMAGENET4437	ImageNet	4,437	Object
IMAGENET8201	ImageNet	8,201	Object
IMAGENET12988	ImageNet	12,988	Object
IMAGENET4000	ImageNet	4,000	Object

2. System description

[Step. 2]

Calculate a score for each concept using visual features.

We selected at most 10 frames from each shot at regular intervals.

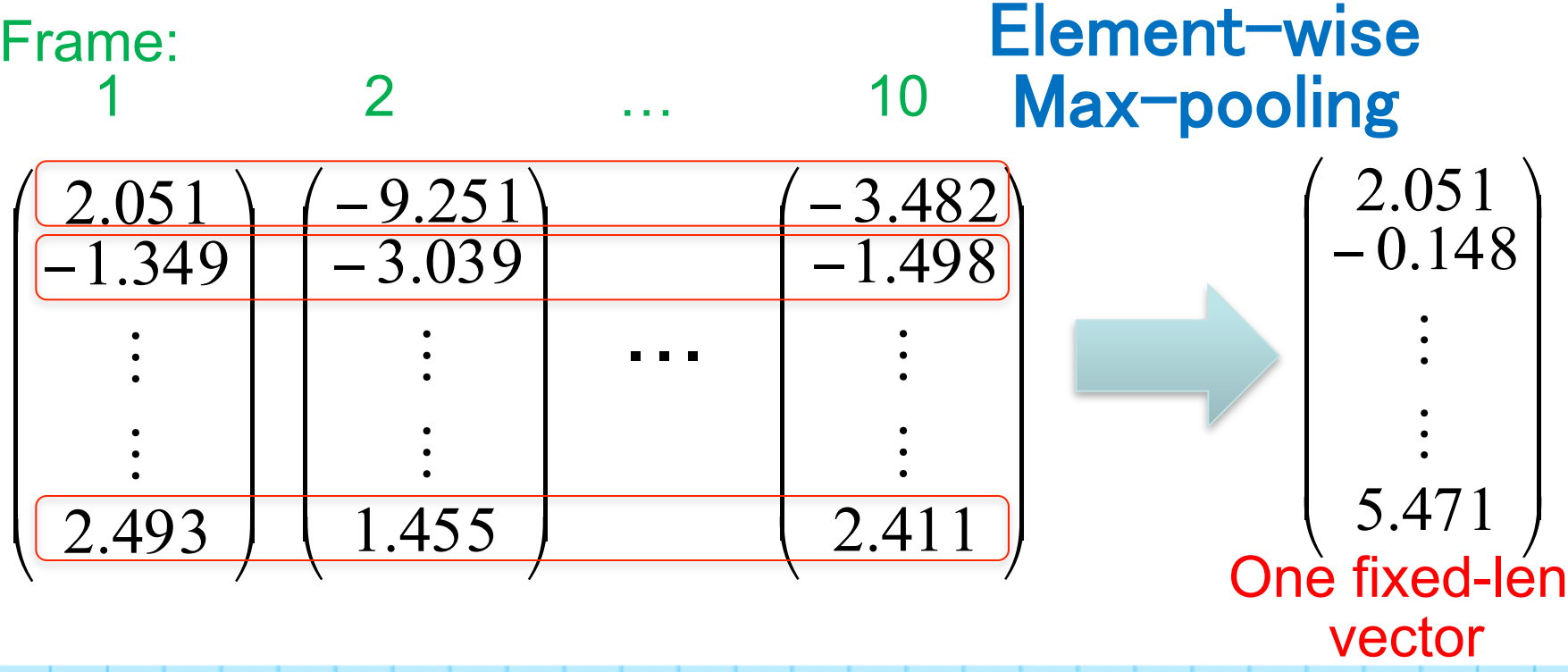


2. System description

[Step. 2]

Calculate a score for each concept using visual features.

Feature vectors were bound to one feature vector by element-wise max-pooling.



2. System description

[Step. 2]

Calculate a score for each concept using visual features.

TRECVID346

- Extract 1024-dimensional features from pool5 layers of pre-trained GoogLeNet model. (trained with ImageNet)
- Train support vector machines (SVMs) for each concept.
- The shot score for each concept was calculated as the distance to hyperplane in the SVM model.

2. System description

[Step. 2]

Calculate a score for each concept using visual features.

PLACES205

- Places205-AlexNet
  (205 scene categories with 2.5 million images)

PLACES365

- Places365-AlexNet
  (365 scene categories with 1.8 million images)

Hybrid1183

- Hybrid-AlexNet
  (205 scene + 978 object categories with 3.6 million images)

provided by MIT

[B. Zhou, 2014] "Learning deep features for scene recognition using places database"

Shot scores were obtained directly from the output layer (before softmax is applied) of the CNNs.

2. System description

[Step. 2]

Calculate a score for each concept using visual features.

ImageNet1000

- AlexNet
  (ImageNet: 1000 object categories)

ImageNet4437, ImageNet8201, ImageNet12988, ImageNet4000

- GoogleNet provided by Univ. of Amsterdam
  (ImageNet: 4437, 8201, 12988, 4000 categories)

[P. Mettes, 2016] "Reorganized Pre-training for Video Event Detection"

Shot scores were obtained directly from the output layer (before softmax is applied) of the CNNs.

2. System description

[Step. 2]

Calculate a score for each concept using visual features.

Score normalization

The score for each semantic concept was normalized over all the test shots such that the maximum and the minimum scores were 1.0 (most probable) and 0.0 (least probable).

Concept selection

No concept name matching a given search keyword.

⇨ Semantically similar concept was chosen by word2vec.

Search keyword did not have a semantically similar concept.

⇨ This keyword was not used.

2. System description

[Step. 3]

Combine the semantic concepts to get the final scores.

Score fusion

Calculate the final scores by score-level fusion

or operator

"walking" or "bicycling"

0.40 0.10 → maximum score → 0.40

and operator

"fountain" and "outdoor"

0.90 0.80

summing score → 0.90 + 0.80 = 1.70

(*) depend on runs

multiplying score → 0.90 x 0.80 = 0.72

3. Submission

3. Submission

Waseda1 run

Total score was simply calculated by multiplying the scores of the selected concepts.

$$\prod_{i=1}^{N} s_i$$

N: # selected concepts

s_i: normalized score

"fountain" and "outdoor"

shot A:	0.70	x	0.10	=	0.07
shot B:	0.40	x	0.30	=	0.12
	⋮		⋮		⋮

Shots having all the selected concepts will tend to appear in the higher ranks.

3. Submission

Waseda2 run

Almost the same as Waseda1 run except for the incorporation of a fusion weight.

$$\prod_{i=1}^{N} s_i^{w_i}$$

fusion weight (= IDF values) calculated from the Microsoft COCO database.

A rare keyword is of higher importance than an ordinary keyword.

	“man”	and	“bookcase”
shot A:	$(0.90)^{1.97}$	x	$(0.70)^{8.23}$
	= 0.81	x	0.05 = 0.04
shot B:	$(0.70)^{1.97}$	x	$(0.90)^{8.23}$
	= 0.50	x	0.42 = 0.21

3. Submission

Waseda3 run

Total score was calculated by summing the scores of the selected concepts.

$$\sum_{i=1}^{N} s_i$$

	“fountain”		and “outdoor”		
shot A:	0.70	+	0.10	=	0.80
shot B:	0.40	+	0.30	=	0.70
	⋮		⋮		⋮

Somewhat looser conditions than multiplying (Waseda1, Waseda2 runs)

3. Submission

Waseda4 run

Similar to Waseda3 except that fusion weight is used.

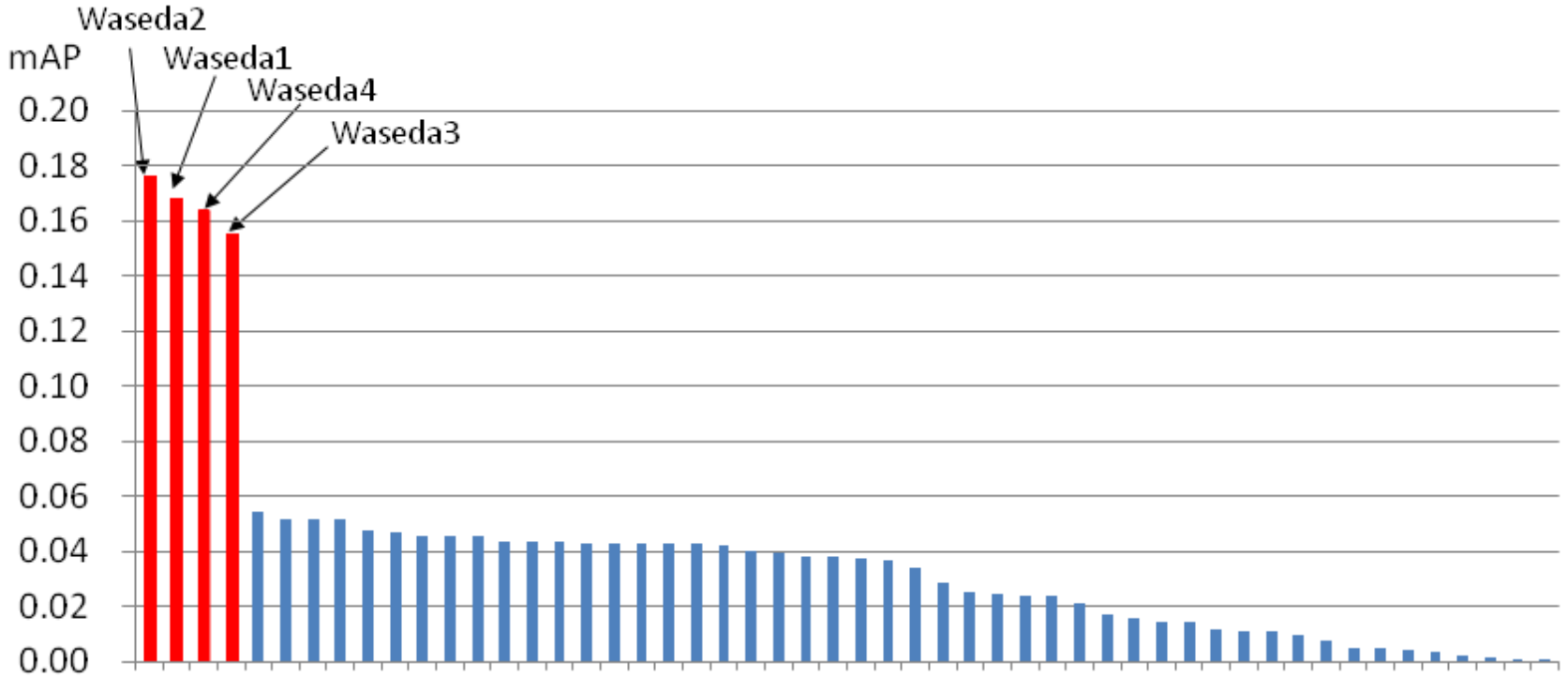
$$\sum_{i=1}^{N} w_i \cdot s_i$$

“man” and “bookcase”

shot A:	(1.97 x 0.90)	+	(8.23 x 0.70)	=	7.53
shot B:	(1.97 x 0.70)	+	(8.23 x 0.90)	=	8.79

4. Results

4. Results



Comparison of Waseda runs with the runs of other teams on IACC_3

Our 2016 submissions ranked between 1 and 4 in a total of 52 runs. Our best run was a mean average precision of 17.7%.

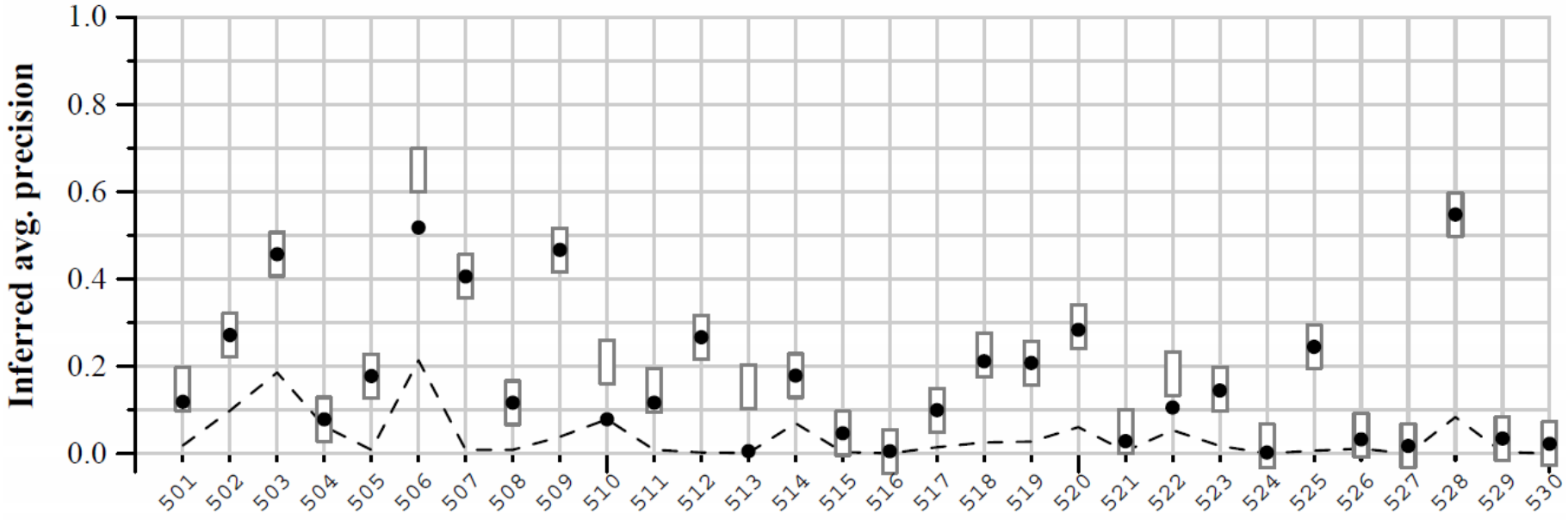
4. Results

Comparison of Waseda runs

Name	Fusion method	Fusion weight	mAP
Waseda1	Multiplying scores		16.9
Waseda2	Multiplying scores	✓	17.7
Waseda3	Summing scores		15.6
Waseda4	Summing scores	✓	16.4

- The stricter condition in which all the concepts in a query phrase must be included has the better performance.
- The rarely seen concepts are much more important for the video retrieval task.

4. Results



Average precision of our best run (Waseda2) for each query.
Run score (dot), median (dashes), and best (box) by query.

The performance was extremely bad for some query phrases.

5. Summary & future works

5. Summary and future works

- We solved the problem of ad-hoc video search by a combination of many semantic concepts.
- We achieved the best performance among all the submission; however, the performance was still relatively low.

Future works

- Increasing the number of semantic concepts, especially those related to action.
- Selecting visually informative keywords.
- Resolving word-sense ambiguities.
- Developing the fully automatic video retrieval system.

Thank you for your attention.

Any questions?