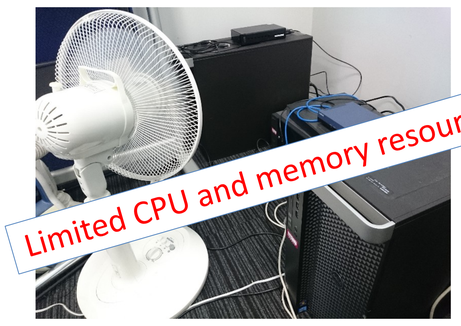
Waseda at TRECVID 2015
Semantic Indexing

Kazuya UEKI, Tetsunori KOBAYASHI (Waseda University)

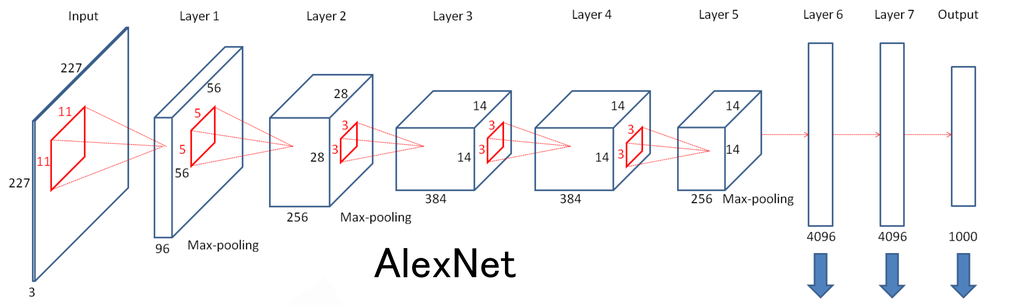
1. System Description

Our computing environment



Two off-the-shelf computers
+ GPUs (Titan Black)

We decided to focus on extracting features only from CNNs.



SIFT, HOG, and etc

Dense trajectories

Instead of using local features or motion features,
6 different CNNs were used.

Score-level fusion

Pipeline

[Step 1] Feature extracting with CNNs

↓

[Step 2] Feature pooling

↓

[Step 3] Classification with SVMs

↓

[Step 4] Classifier fusion

[Step 1] Feature extraction with CNNs

(1) ImageNet

- Trained with the ImageNet dataset (1.2 million images and 1,000 categories)
- Provided with the Caffe (CNN) library

(2) Finetune

- Created by finetuning ImageNet model for TRECVID SIN task
- 1 million keyframe images
- 346 concepts (# of units in the output layer: 346)

(3) Gradient

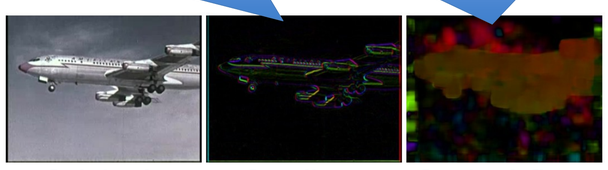
- Substitute edge features with CNN features
- Train with 1 million gradient images (346 concepts)

Color: Orientation of gradient
Brightness: Magnitude of the orientation gradients

(4) OpticalFlow

- Substitute motion features with CNN features
- Train with 1 million optical flow images (346 concepts)

Color: Orientation of the optical flow
Brightness: Magnitude of the optical flow



Original images | Gradient images | Optical flow images

(5) Places

- Scene recognition model
- Trained on 205 scene categories
- 2.5 million images
- Provided by MIT (Caffe model zoo) [※]

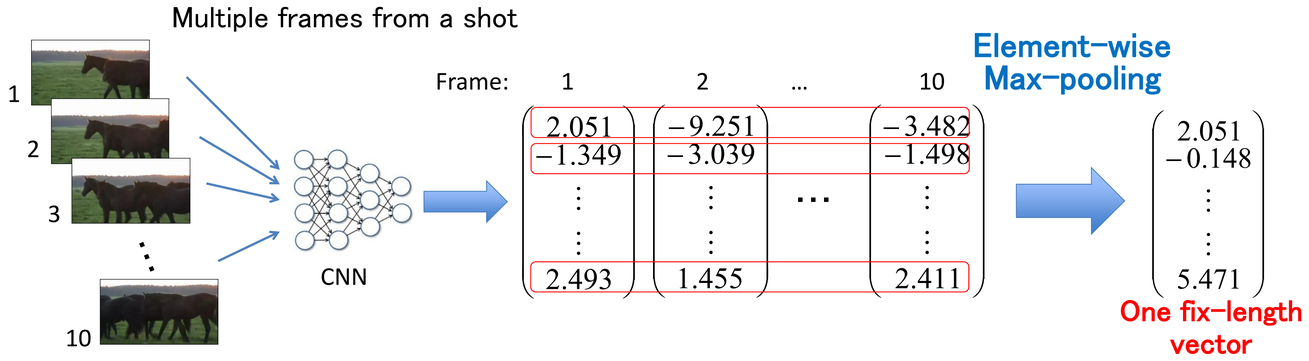
(6) Hybrid

- Scene and object recognition model
- Trained on 1,183 categories (205 scene categories + 978 object categories)
- 3.6 million images
- Provided by MIT (Caffe model zoo) [※]

[※] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. "Learning Deep Features for Scene Recognition using Places Database." Advances in Neural Information Processing Systems 27 (NIPS), 2014.

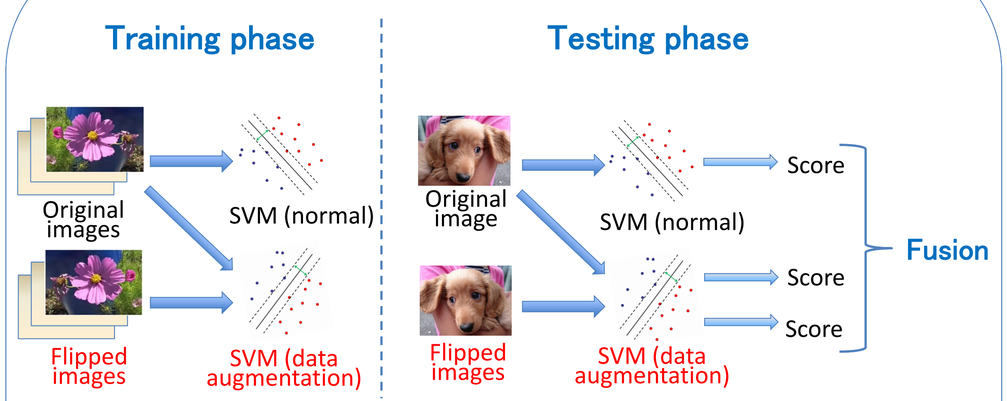
[Step 2] Feature pooling

Multiple frames from a shot



We selected a maximum of 10 frames from a shot at regular intervals.

[Step 3] Classification with SVMs



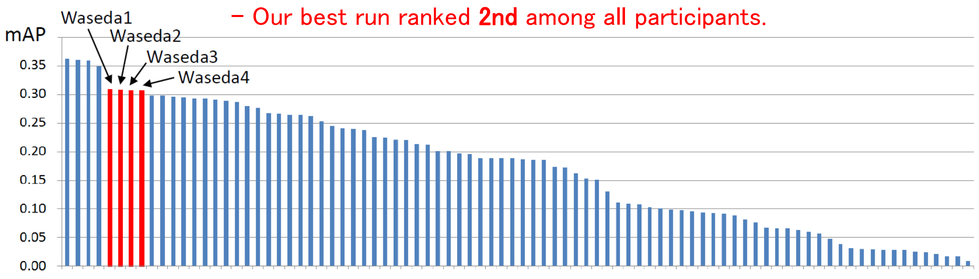
Scores from the following 3 scores were combined.
- Original images used for both training and testing
- Both original and flipped images used for training, but only original images used for testing
- Both original and flipped images used for training, and only flipped images used for testing.

[Step 4] Classifier fusion

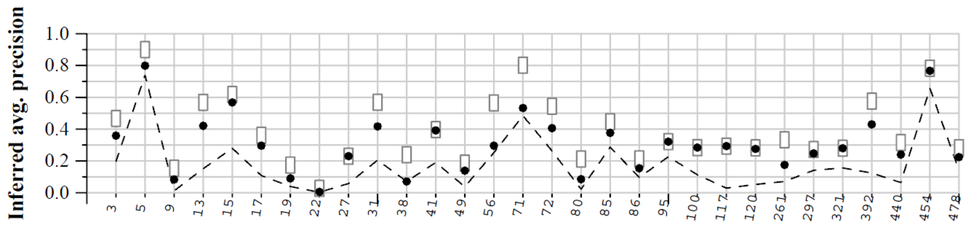
- Waseda4: Fusion weight of 2 for **ImageNet**, **Finetune**, **Places** and **Hybrid** models. Fusion weight of 1 for **Hybrid** and **Gradient** models.
- Waseda3: Fusion weight were optimized to improve the mAP of 30 concepts.
- Waseda2: Fusion weight were optimized to improve the mAP of 60 concepts.
- Waseda1: Fusion weight were optimized to improve the average precision of each concept.

2. Results of Submitted Runs

- Our 2015 submissions ranked between 5 and 8 in a total of 86 runs.
- Our best run ranked **2nd** among all participants.



Comparison of Waseda runs with the runs of other teams on IACC 2 C.



Average precision of our best run (Waseda1) for each semantic concept.

One of our runs achieved the best average precision for some concepts: "Cheering", "Demonstration Or Protest", "Press Conference", "Running", "Telephones", "Throwing", and "Lakes".

The mAPs for individual models with the TRECVID 2015 SIN testing set.

Model	Layer	Train: Original image	Train: Original + Flipped images	
		Test: Original images	Test: Original images	Test: Flipped images
ImageNet	6	24.02	24.14	23.75
	7	23.61	23.89	23.53
	8	18. 82 (*1)	19.08 (*1)	18.70 (*1)
Finetune	6	23.50	23.80	23.84
	7	23.29	23.39	23.44
	8	21.53	21.90	21.78
Gradient	6	20.74	19.41	19.03
	7	19.82	18.95	19.17
	8	17.71	17.26	17.35
Optical Flow	6	14.21	14.43	13.99
	7	13.22	13.34	13.42
	8	13.12	13.43	13.56
Places	6	23.40	23.61	23.74
	7	22.29	22.41	22.20
	8	--- (*2)	--- (*2)	--- (*2)
Hybrid	6	25.12	24.75	24.34
	7	25.52	25.17	24.79
	8	23.20	22.93	22.88

(*1) This result includes some errors. After rectifying the errors, the mAPs were changed to 22.04, 22.20, and 21.74, respectively

(*2) We could not finish the calculation by the submission deadline. After the submission, we evaluated the performance and found that the mAPs were 19.73, 19.86, and 19.59, respectively.

Classifier fusion

Waseda1: **30.86**
Waseda2: 30.73
Waseda3: 30.69
Waseda4: 30.69

Conclusion

- Despite the simplicity of our method, it achieved relatively high performance.
- The performance of semantic video indexing was still extremely low.
- In the future, we will investigate the root causes of this poor performance and evaluate the options for improving it.