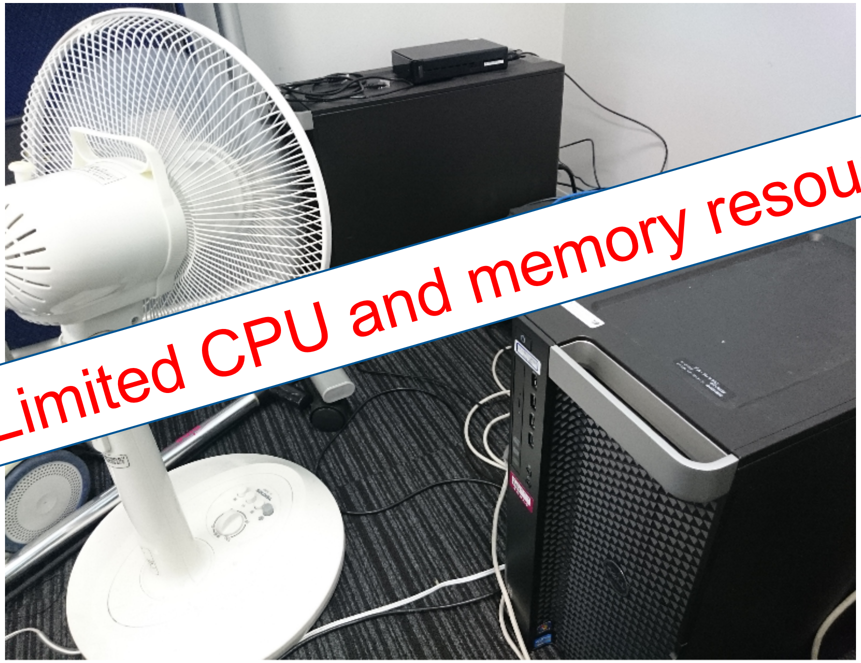
# Waseda at TRECVID 2015

## Semantic Indexing (SIN)

Kazuya UEKI and Tetsunori KOBAYASHI
Waseda University

# 1. System Description

## Our computing environment

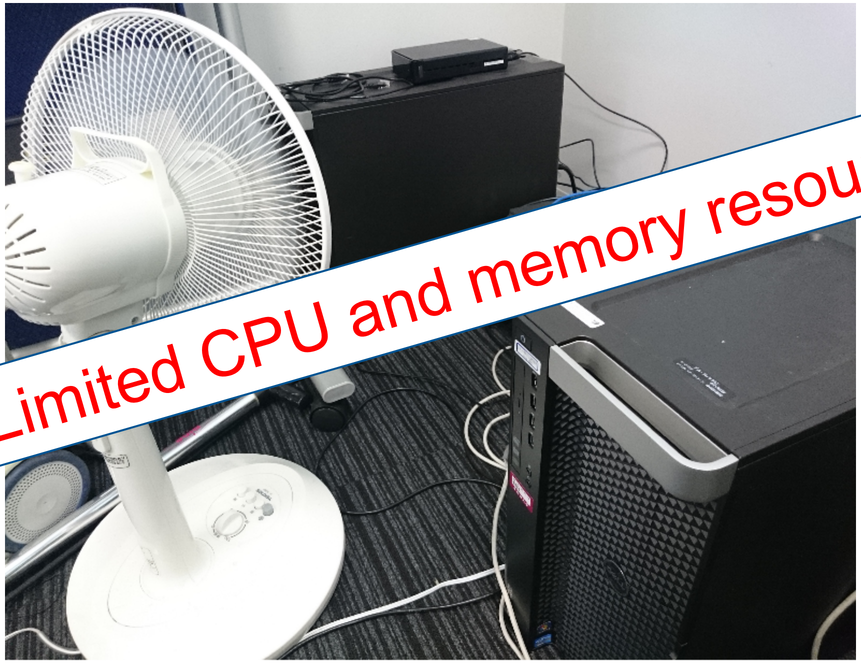


Limited CPU and memory resources

Two off-the-shelf computers.

# 1. System description

## Our computing environment



Limited CPU and memory resources

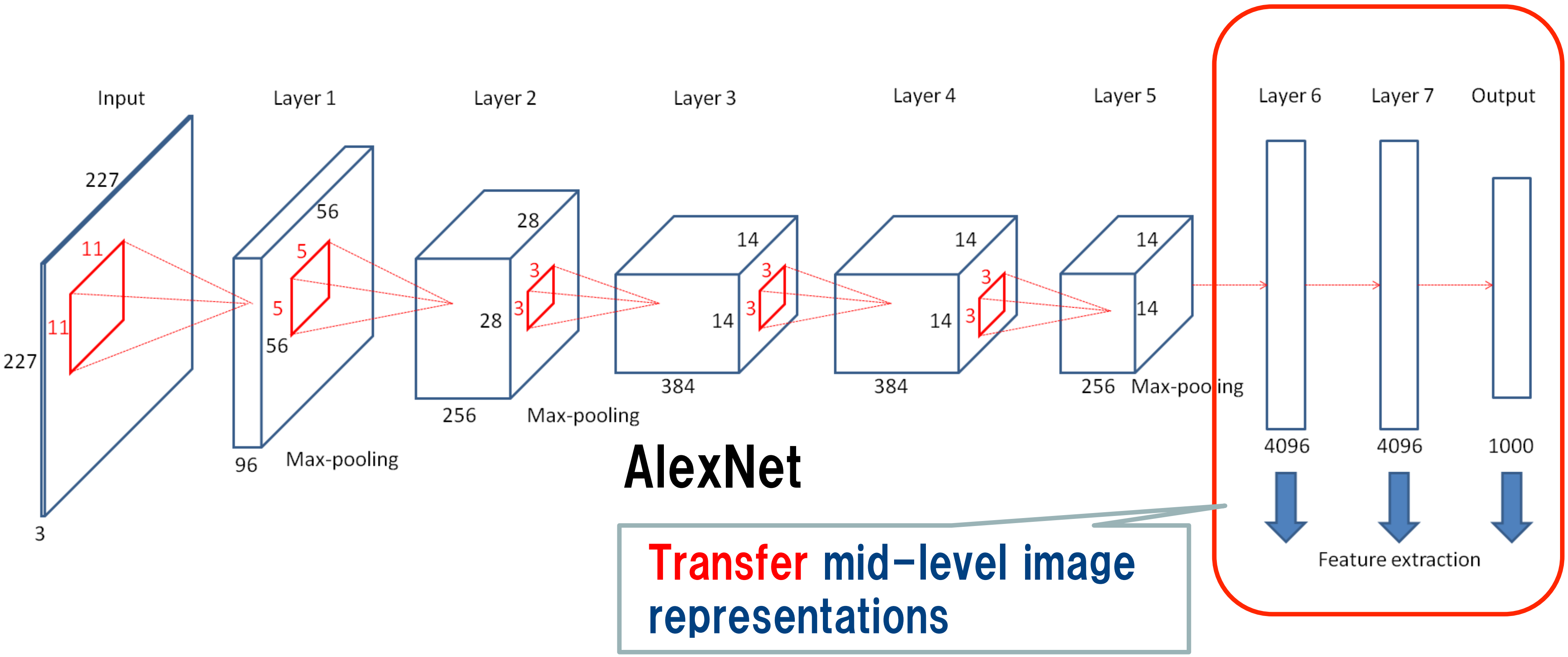
**Two off-the-shelf computers.**

**But these computers each have two GPUs.**

(Titan Black)

# 1. System description

For this year's submission, we decided to focus on extracting features **only from CNNs**.




We did not use local features (SIFT or HOG), motion features (dense trajectories), and audio features.

# 1. System description

## Semantic indexing pipelines:

[Step 1] Feature extraction using multiple CNNs

↓

[Step 2] Feature pooling

↓

[Step 3] Classification with SVMs

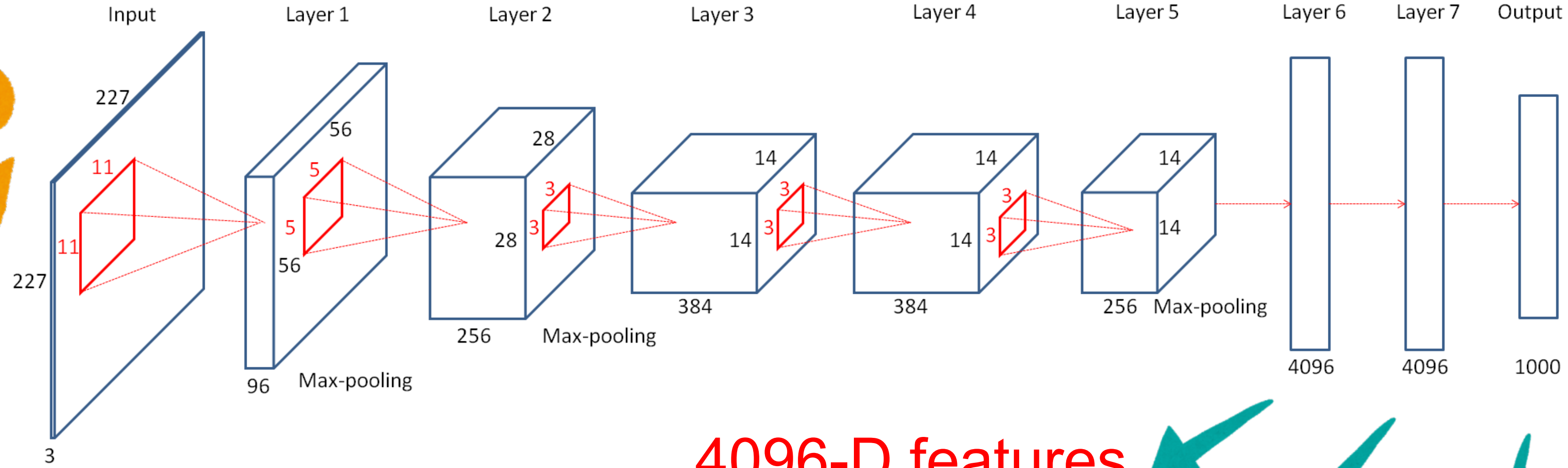
↓

[Step 4] Fusion of multiple score outputs

# 1. System description

## [Step 1] Feature extraction using multiple CNNs

The network structure: AlexNet



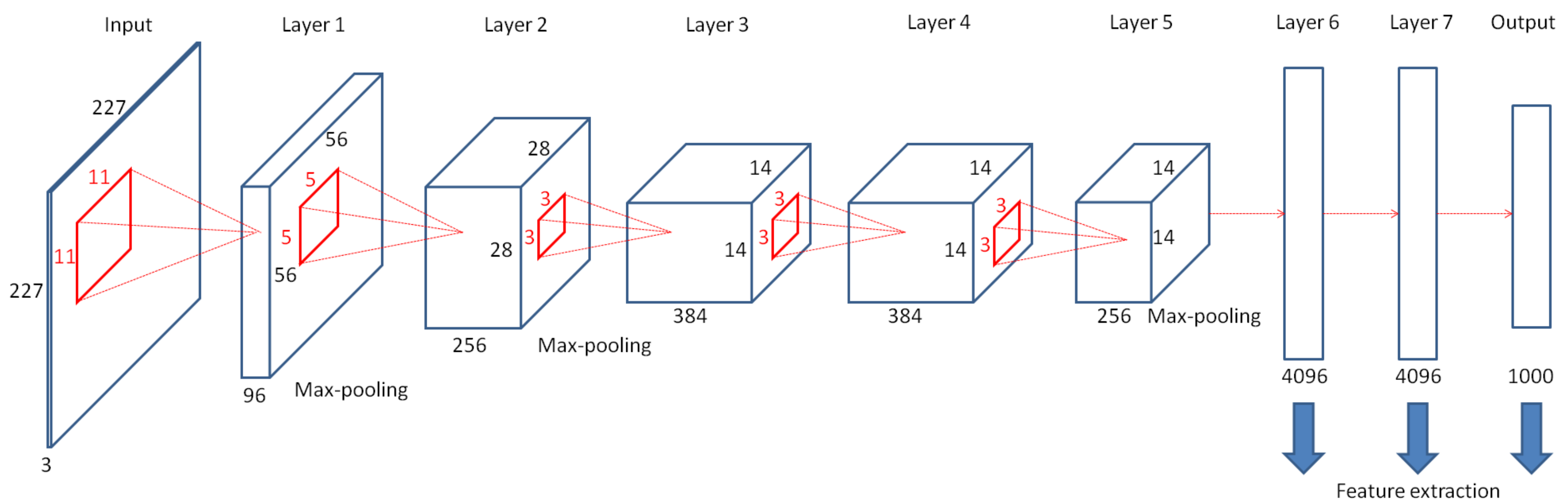

4096-D features

4096-D features

1000-D features

# 1. System description

## [Step 1] Feature extraction using multiple CNNs




SIFT, HOG, and etc

Dense trajectories

**Instead of using <u>local features</u> or <u>motion features</u>,**
**6 different CNNs were used.**

# 1. System description

[Step 1] Feature extraction using multiple 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)

# 1. System description

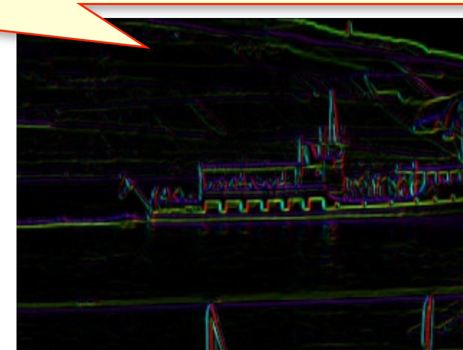
## [Step 1] Feature extraction using multiple CNNs

### (3) Gradient

- Substitute edge features with CNN features
- Trained with 1 million gradient images
- 346 concepts

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

Original image



Gradient image

## [Step 1] Feature extraction using multiple CNNs

### (4) OpticalFlow

- Substitute **motion features** with CNN features
- Trained with 1 million **optical flow images**
- 346 concepts

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

Original image

Optical flow image

# 1. System description

[Step 1] Feature extraction using multiple CNNs

## (5) Places

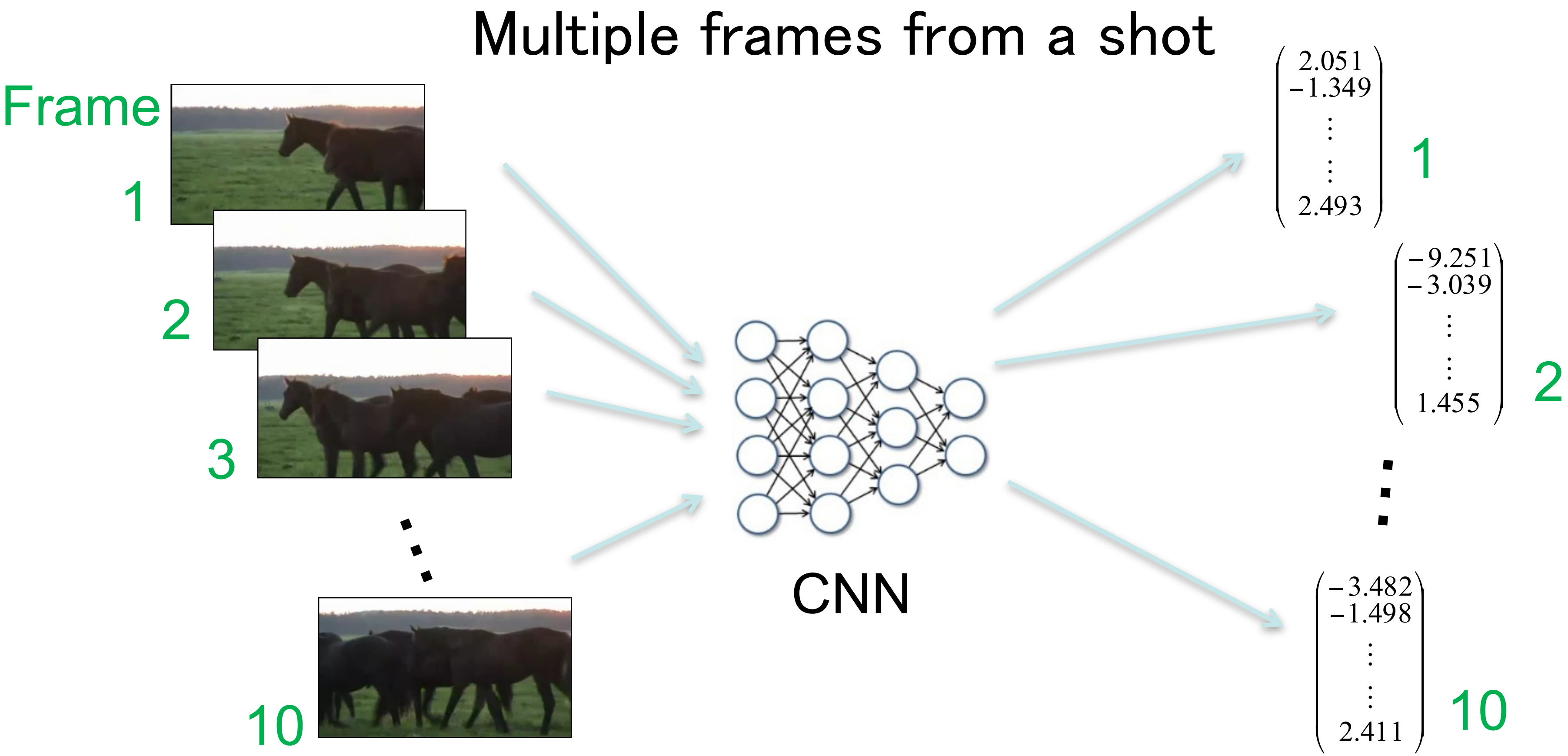
- Scene recognition model
- Trained on 205 scene categories
- 2.5 million images
- Provided by MIT (Caffe model zoo)

## (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)

# 1. System description

## [Step 2] Feature pooling




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

# 1. System description

## [Step 2] Feature pooling

Frame:

$$
\begin{array}{cccc}
1 & 2 & \cdots & 10 \\
\begin{pmatrix} 2.051 \\ -1.349 \\ \vdots \\ \vdots \\ 2.493 \end{pmatrix} &
\begin{pmatrix} -9.251 \\ -3.039 \\ \vdots \\ \vdots \\ 1.455 \end{pmatrix} &
\therefore &
\begin{pmatrix} -3.482 \\ -1.498 \\ \vdots \\ \vdots \\ 2.411 \end{pmatrix}
\end{array}
\Rightarrow
\begin{pmatrix} 2.051 \\ -0.148 \\ \vdots \\ \vdots \\ 5.471 \end{pmatrix}
$$

**Element-wise Max-pooling**

One fixed-length vector

**The values of the elements in the same dimension were compared across 10 sets, and the maximum value was selected.**

# 1. System description

[Step 3] Classification with SVMs

6th, 7th and 8th layer

- Create a separate SVM using features from <u>each layer</u>

  To reduce the computational cost and the toll on memory resources

- Use approximately 20,000-30,000 shots for each concepts

  With roughly the same number of positive and negative samples

- Utilize flipped images during both the training and the testing

  To enrich the variations of the training and the testing sets

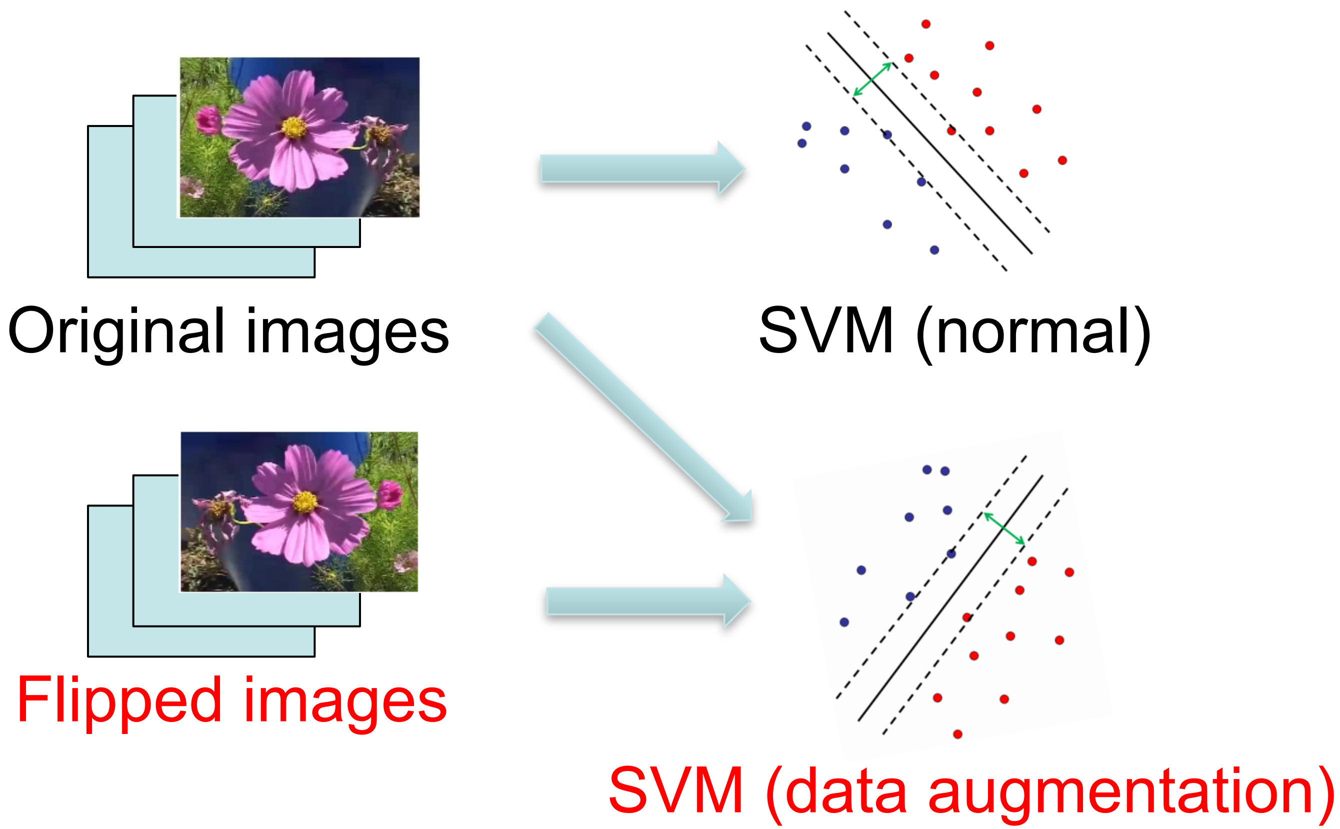
  There are far fewer positive samples than negative samples.

  ⇨ Use the flipped images exclusively for the positive samples.

# 1. System description

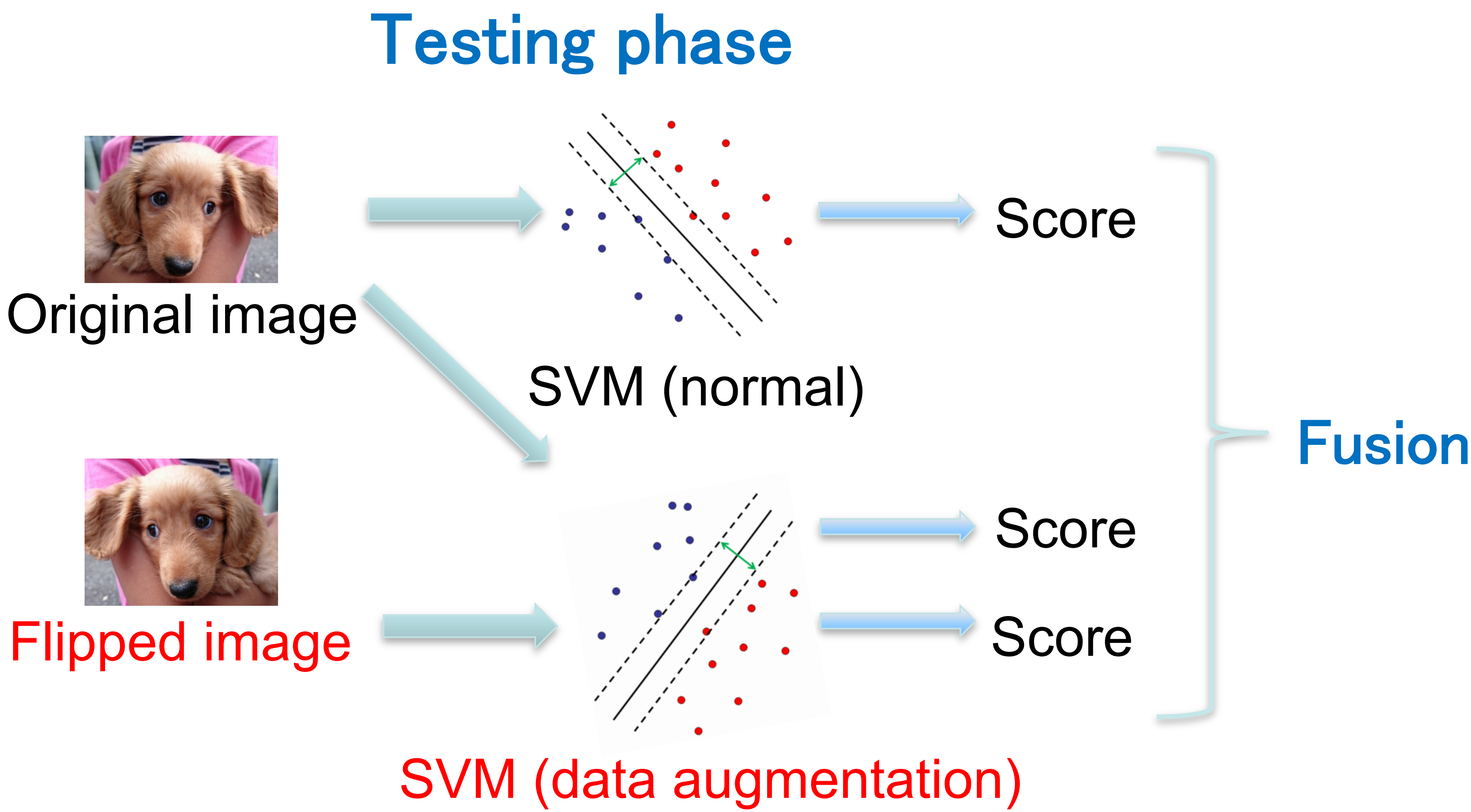
## [Step 3] Classification with SVMs

### Training phase



# 1. System description

## [Step 3] Classification with SVMs



## [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] Fusion of multiple score outputs

- 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.

⇒ Fusion weight optimization did not offer significant improvements over averaging of scores.

## Submission results

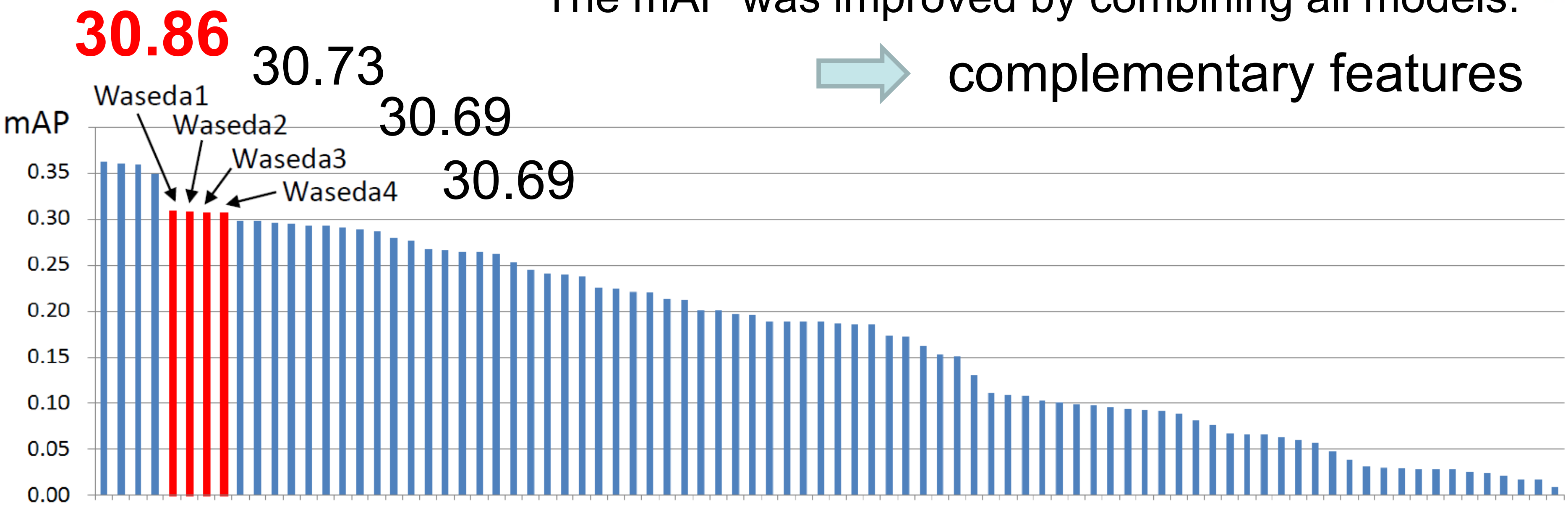
The mAPs for individual models with the TRECVID 2015 SIN testing set.

| Model | Layer | Train: original images | 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 |
| OpticalFlow | 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 |

# 2. Results of Submitted Runs

The mAP was improved by combining all models.

⇒ complementary features

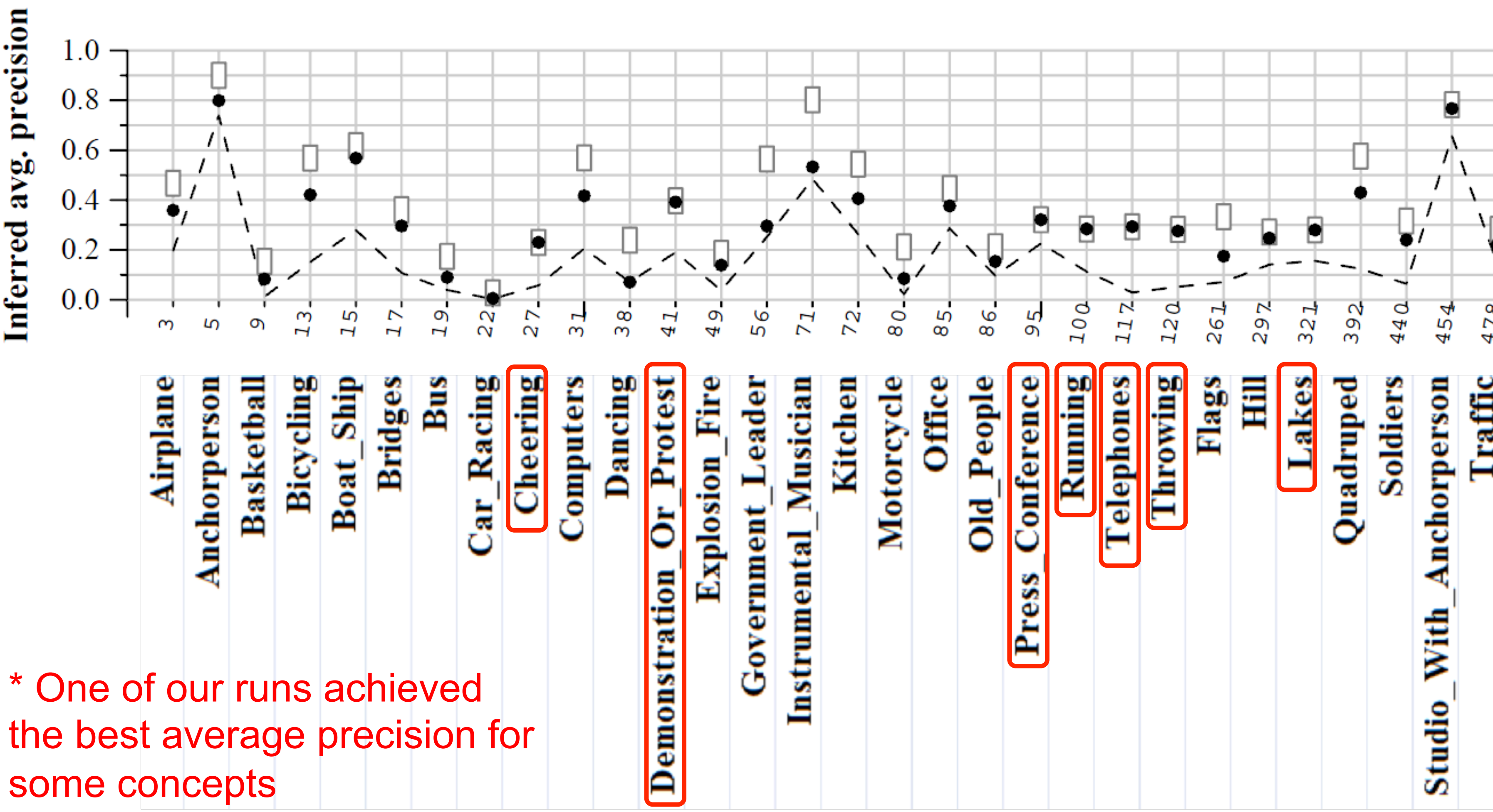



Comparison of Waseda runs with the runs of other teams on IACC 2 C.

- Our 2015 submissions ranked between 5 and 8 in a total of 86 runs.
- Our best run was an mAP of 30.86%, which ranked 2nd among all participants.

# 2. Results of Submitted Runs

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




* One of our runs achieved the best average precision for some concepts

# 3. Summary and future works

- 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.

Thank you for your attention.

Any questions?