Kobe University and Kindai University at TRECVID 2018 AVS Task

Kimiaki Shirahama[1], **He Zhenying**[2] and **Kuniaki Uehara**[2]
Department of Informatics, Kindai University
Graduate School of System Informatics, Kobe University

Abstract

This year we addressed the following two points:

- How to fuse concept detection scores for accurate retrieval
  Cascade-based approach that uses a sequence of stages to gradually filter out irrelevant shots
- How to deal with a topic requiring the number of objects or their relation
  Object detection to analyze detected regions

Concept detection

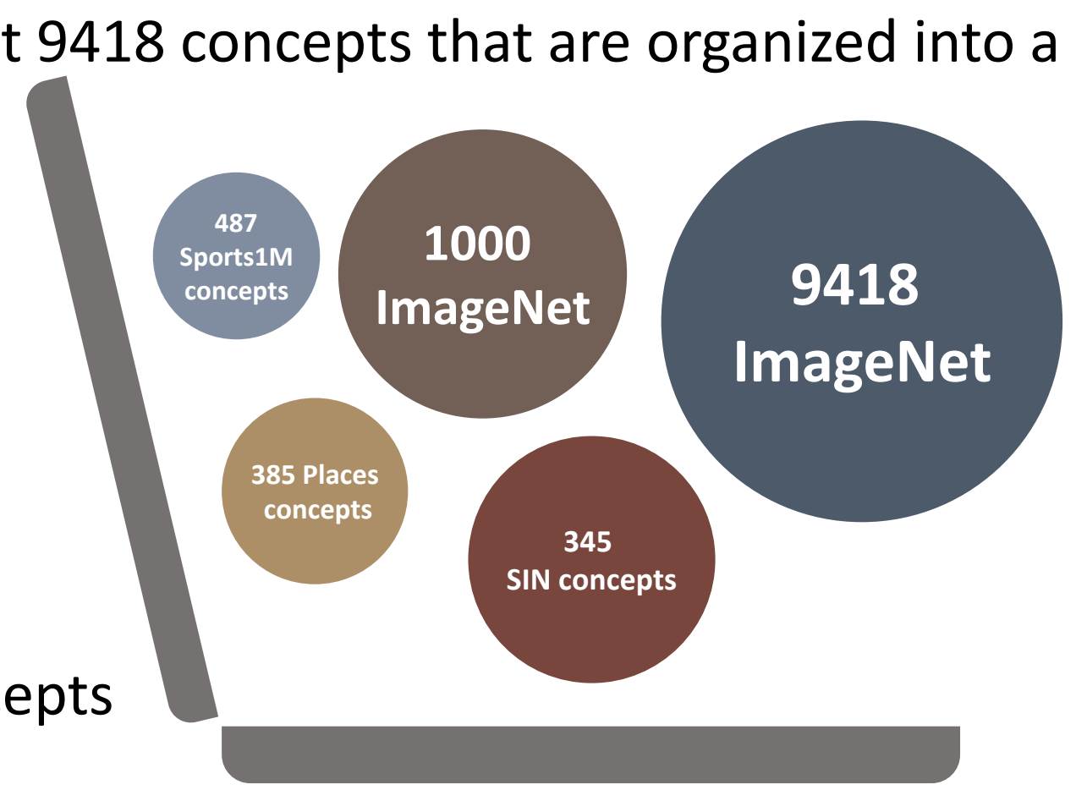
345 SIN concepts: Detection scores that are provided by ITI-CERTH team and obtained by SVM-based fine-tuning of pre-trained network

1000 ImageNet concepts: **ResNet152** implementation in YOLO to detect 1000 concepts in ImageNet

9418 ImageNet concepts: **darknet9000** to detect 9418 concepts that are organized into a hierarchical tree

385 Places concepts: ResNet152 fine-tuned for 365 scene concepts defined in Places365, as well as max-pooling to obtain detection scores for their 20 super-concepts

487 Sports1M concepts: C3D to detect 487 concepts defined in Sports1M dataset.



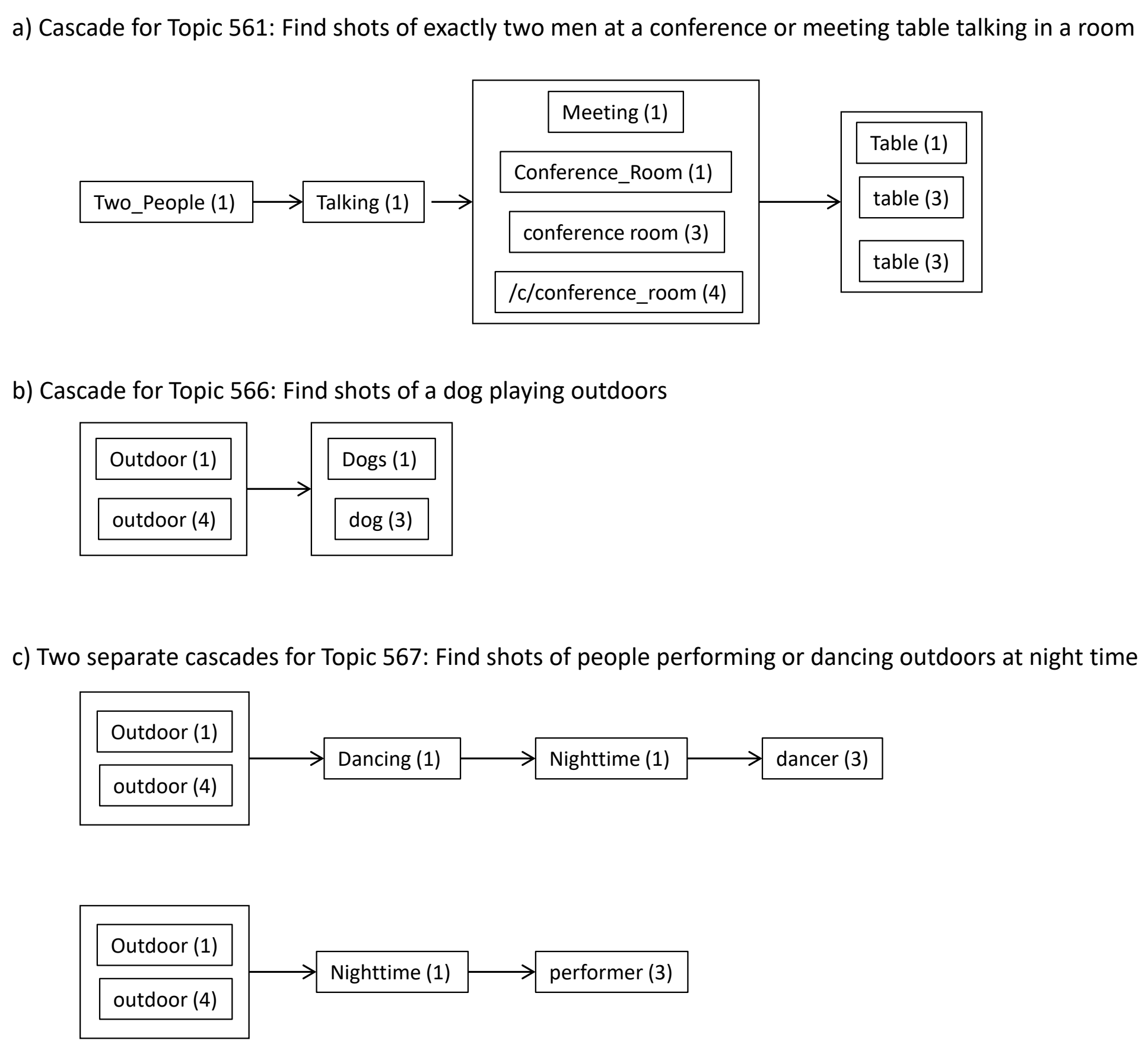
Concept Selection

1. **Generality: Use the most general concept**
   e.g. Topic 566: Find shots of a dog playing outdoors
   Negative concept: indoor
2. **Specificity: Use a specific concept deduced from a phase in a topic**
   e.g. Topic 563: Find shots of one or more people on a moving boat in the water
   boatman

Cascade Construction

Selected concepts are organized into a cascade where each concept is associated with one **stage**

a) Order of stage: As a concept is **more general**, the corresponding stage is placed **earlier**
b) Parallel: Multiple concepts representing the same (or very similar) meaning are placed in parallel
c) Separate cascades: Multiple cascades are used for a topic including "or"


a) Cascade for Topic 561: Find shots of exactly two men at a conference or meeting table talking in a room

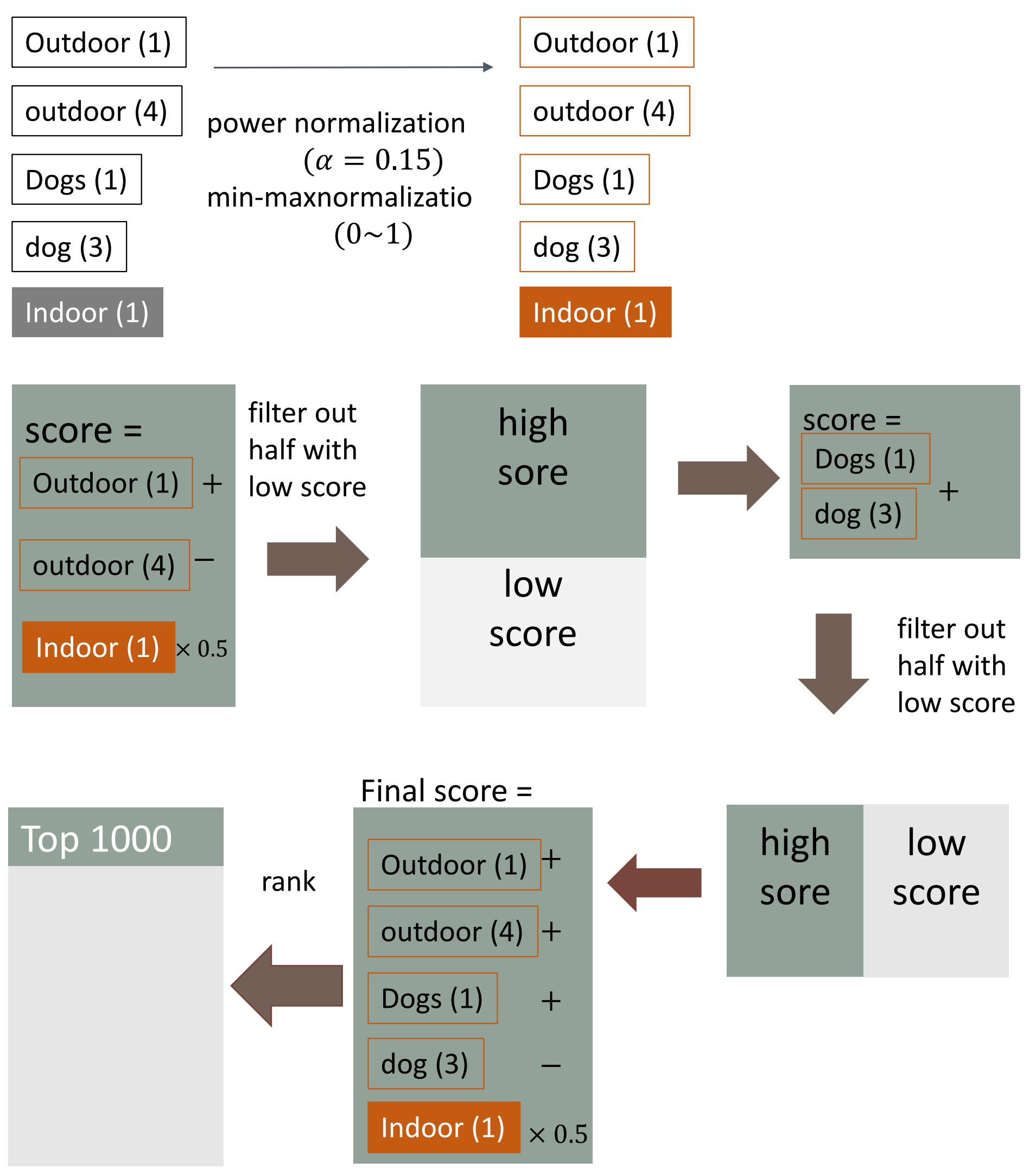
b) Cascade for Topic 566: Find shots of a dog playing outdoors

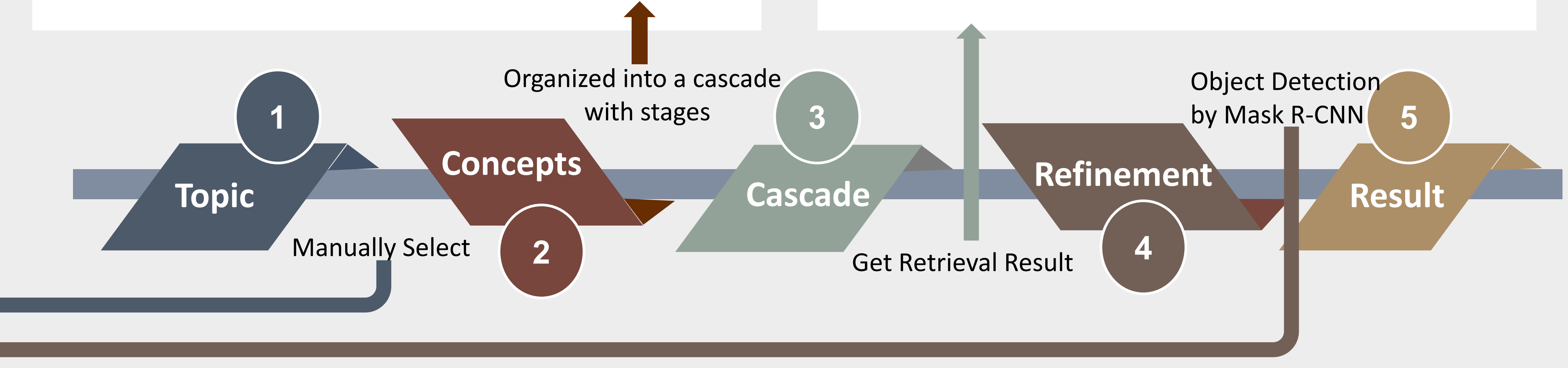
c) Two separate cascades for Topic 567: Find shots of people performing or dancing outdoors at night time

Cascade-based Retrieval

1. Normalize detection scores
2. Filter out
3. Rank

Topic 566: Find shots of a dog playing outdoors





Refinement by Object Detection

Topics with requirement on the number of objects or their spatial relationship

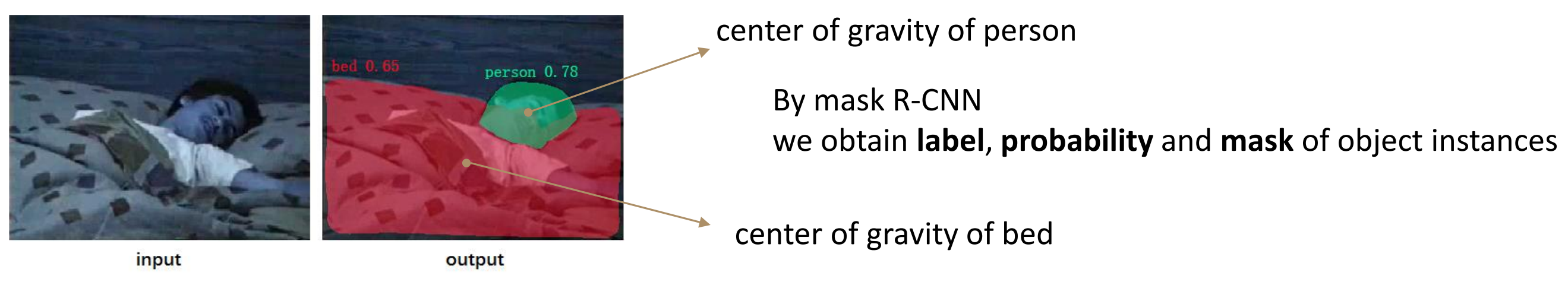
Topic 561: Find shots of exactly **two** men at a conference or meeting table talking in a room.

Number of object

Topic 584: Find shots of a person **lying on** a bed.

Spatial relationship between objects

Object detection by Mask R-CNN

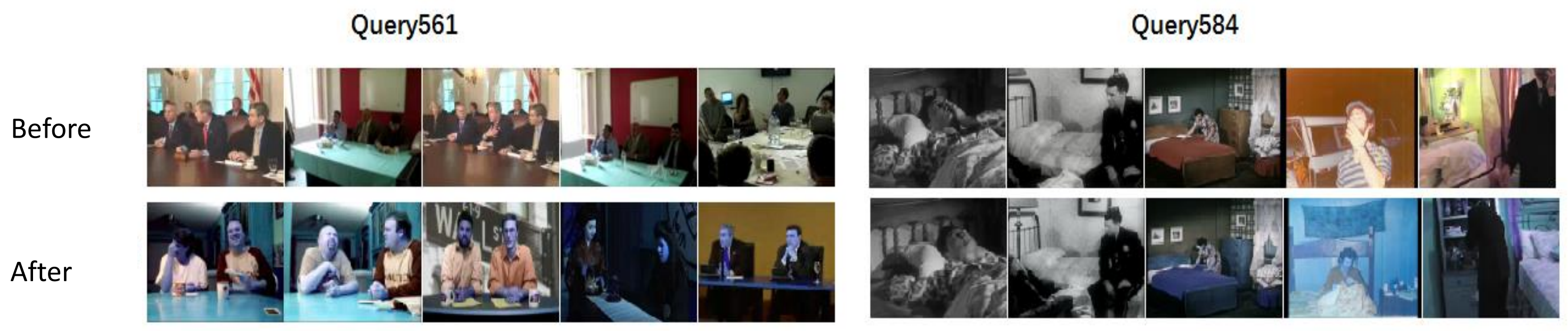


By mask R-CNN
we obtain **label**, **probability** and **mask** of object instances

Examination of the top 10000 shots retrieved by the cascade-based approach

- ❑ **Number of objects:**
  - Use the number of instances with the same label in a keyframe
- ❑ **Spatial relationship between objects:**
  - Get center of gravity of each object instance by calculating average of all pixel coordinate in the instance mask
  - Determine the spatial relationship by comparing the center of gravity of an object to the one of another object

Shot filtering



Filter out shots with not exactly two people(works well)

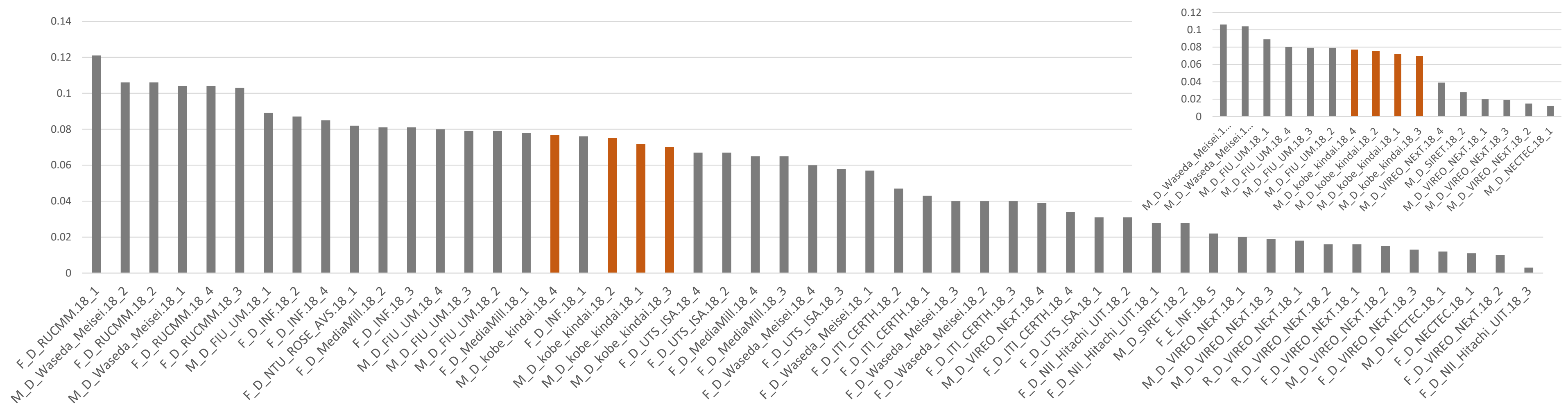
Filter out shots in which person's position is lower than bed's(doesn't work well)

Future work

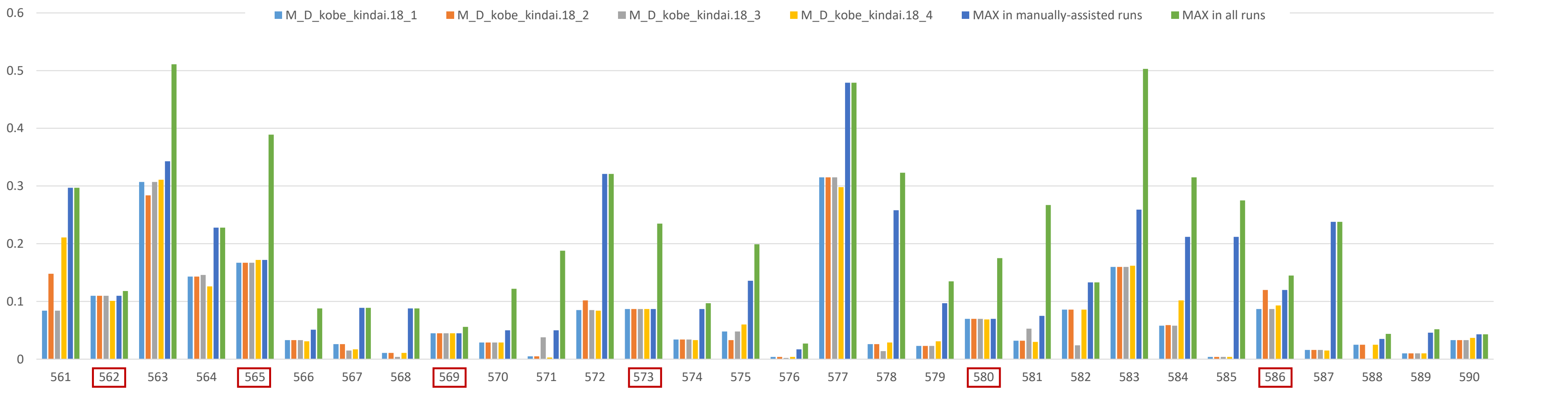
- Adopt an "embedding-based" approach to avoid cumbersome issues in the concept-based approach, like concept selection and score fusion/pooling
- Use Deep relational network to specifically predict the complex relationships between objects.

Results

- **M_D_kobe_kindai.18_1:Baseline** that uses the **cascade-based approach** without object detection.
- **M_D_kobe_kindai.18_2:** Refinement of shots retrieved by M_D_kobe_kindai.18_1 with **object detection**
- **M_D_kobe_kindai.18_3: Slightly different sets of concepts** from M_D_kobe_kindai.18_1 for some topics
- **M_D_kobe_kindai.18_4: Simple summation of detection scores** for the selected concepts.



1. M_D_kobe_kindai.18 4 is ranked at the seventh place among 16 runs in the manually-assisted category. (Our team is ranked at the third place among six teams)
2. Adoption of the **large concept vocabulary** leads to good performances.

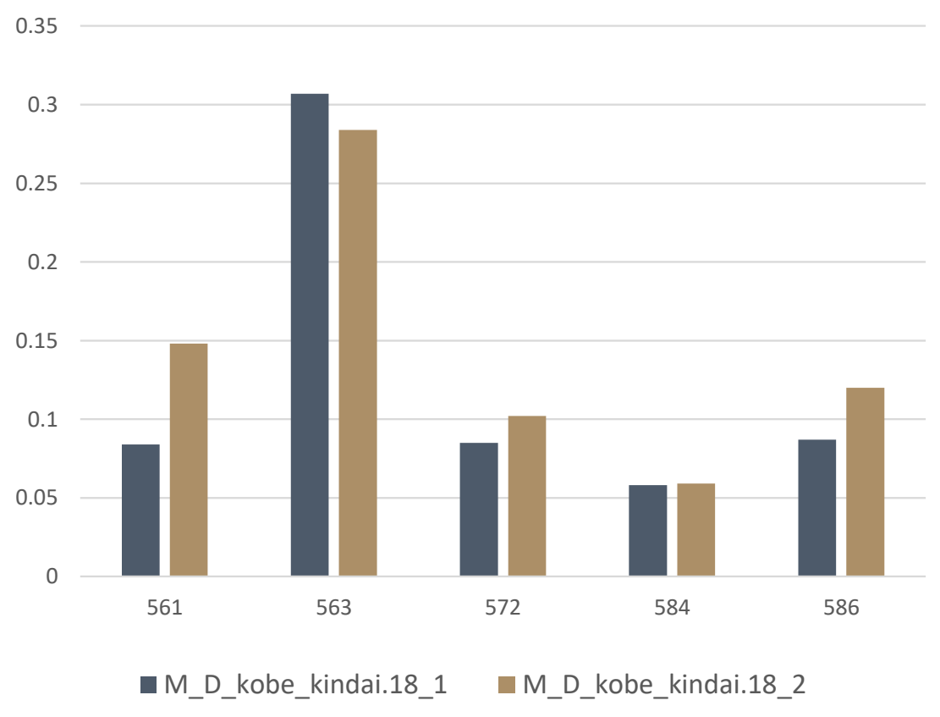


1. Our runs achieved the best average precisions for the six topics in the manually-assisted category.
2. No significant difference is observed between using the cascade-based approach and not-using it.
3. Cascade-based approach **reduces search time** (from 5.9s to 4.0s).

Topic 584: a person lying on a bed

Topic 569: people standing in line outdoors.

For **complex relations** between objects like waving flags and pouring liquid, our current method only considers their co-occurrence



For topics requiring spatial relationship

- The person's center of gravity is higher than bed's, but the person is sitting on the bed.

For complex topic

- Although we obtained masks of object instances, it's difficult to define which situation is correct.)

1. **Object detection** effectively refines retrieval results, especially when the number of objects is required.
2. For topics requiring **spatial relationship**, object detection **didn't work** as **good** as we expected.
3. For complex shot, it's difficult to define which shot is correct