ECCV 2018 Tutorial

Video Recognition and Retrieval at the TRECVID Benchmark

Lecture 3: Ad-hoc Video Search (AVS) Task

Kazuya UEKI

Waseda University

Meisei University

Tutorial Outline

Part I: the Ad-hoc Video Search (AVS) task

- Goal of the AVS task
- Definition of the AVS task

Part II: Results of submitted systems

- Some participants’ implementations
- Evaluation results

Part III: Summary and future works

Part I:
the Ad-hoc Video Search (AVS) task

What is the AVS task?

Zero-shot Video retrieval using a query phrase

a person holding a poster on the street at daytime

The major difficulty in this task:

- A system must retrieve videos under conditions where no training videos match a query phrase.
- A system have to retrieve video sequences that simultaneously contain multiple detection targets (*concepts*), such as persons, objects, scenes, and actions.

Ad-hoc Video Search Task Definition

- **Goal**: promote progress in content-based retrieval based on end user ad-hoc queries that include persons, objects, locations, activities and their combinations.

• Who : concrete objects and being (kind of persons, animals, things)
• What : are the objects and/or beings doing ?
(generic actions, conditions/state)
• Where : locale, site, place, geographic, architectural
• When : time of day, season

- **Task**: Given a test collection, a query, and a master shot boundary reference, return a ranked list of at most 1,000 shots (out of 335,944) which best satisfy the need.

- **Testing data**: 4,593 Internet Archive videos (IACC.3), 600 total hours with video durations between 6.5 min to 9.5 min.

TRECVID 2017 queries by complexity

Person + Action + Object + Location

- Find shots of one or more people eating food at a table indoors
- Find shots of one or more people driving snowmobiles in the snow
- Find shots of a man sitting down on a couch in a room
- Find shots of a person talking behind a podium wearing a suit outdoors during daytime
- Find shots of a person standing in front of a brick building or wall

Person + Action + Location

- Find shots of children playing in a playground
- Find shots of one or more people swimming in a swimming pool
- Find shots of a crowd of people attending a football game in a stadium
- Find shots of an adult person running in a city street

TRECVID 2017 queries by complexity

Person + Action/state + Object

- Find shots of a person riding a horse including horse-drawn carts
- Find shots of a person wearing any kind of hat
- Find shots of a person talking on a cell phone
- Find shots of a person holding or operating a tv or movie camera
- Find shots of a person holding or opening a briefcase
- Find shots of a person wearing a blue shirt
- Find shots of person holding, throwing or playing with a balloon
- Find shots of person wearing a scaft
- Find shots of a person holding, opening, closing or handing over a box

Person + Action

- Find shots of a person communicating using sign language
- Find shots of a child or group of children dancing
- Find shots of people marching in a parade
- Find shots of a male person falling down

TRECVID 2017 queries by complexity

Person + Object + Location

- Find shots of a man and woman inside a car

Person + Location

- Find shots of a chef or cook in a kitchen
- Find shots of a blond female indoors

Person + Object

- Find shots of a person with a gun visible

Object + Location

- Find shots of a map indoors

Object

- Find shots of vegetables and/or fruits
- Find shots of a newspaper
- Find shots of at least two planes both visible

Training and run types

Four training data types:

- ✓ A – used only IACC training data (0 runs)
- ✓ D – used any other training data (40 runs)
- ✓ E – used only training data collected automatically using only the query text (12 runs)
- ✓ F – used only training data collected automatically using a query built manually from the given query text (0 runs)

Two run submission types:

- ✓Manually-assisted (M) – Query built manually (19 runs)
- ✓Fully automatic (F) – System uses official query directly (33 runs)

Finishers: 10 out of 20

Team	Organization	M	F
INF	Renmin University; Shandong Normal University; Chongqing university of posts and telecommunications; Carnegie Mellon University	-	4
kobe_nict_siegen	Kobe University, Japan Center for Information and Neural Networks, National Institute of Information and Communications Technology (NICT), Japan Pattern Recognition Group, University of Siegen, Germany	3	-
ITI_CERTH	Information Technologies Institute, Centre for Research and Technology Hellas	-	4
ITEC_UNIKLU	Klagenfurt University	4	4
NII_Hitachi_UIT	National Institute of Informatics, Japan (NII); Hitachi, Ltd; University of Information Technology, VNU-HCM, Vietnam (HCM-UIT)	-	4
MediaMill	University of Amsterdam	-	4
Waseda_Meisei	Waseda University; Meisei University	4	4
VIREO	City University of Hong Kong	4	4
EURECOM	EURECOM	-	4
FIU_UM	Florida International University, University of Miami	4	-

Evaluation

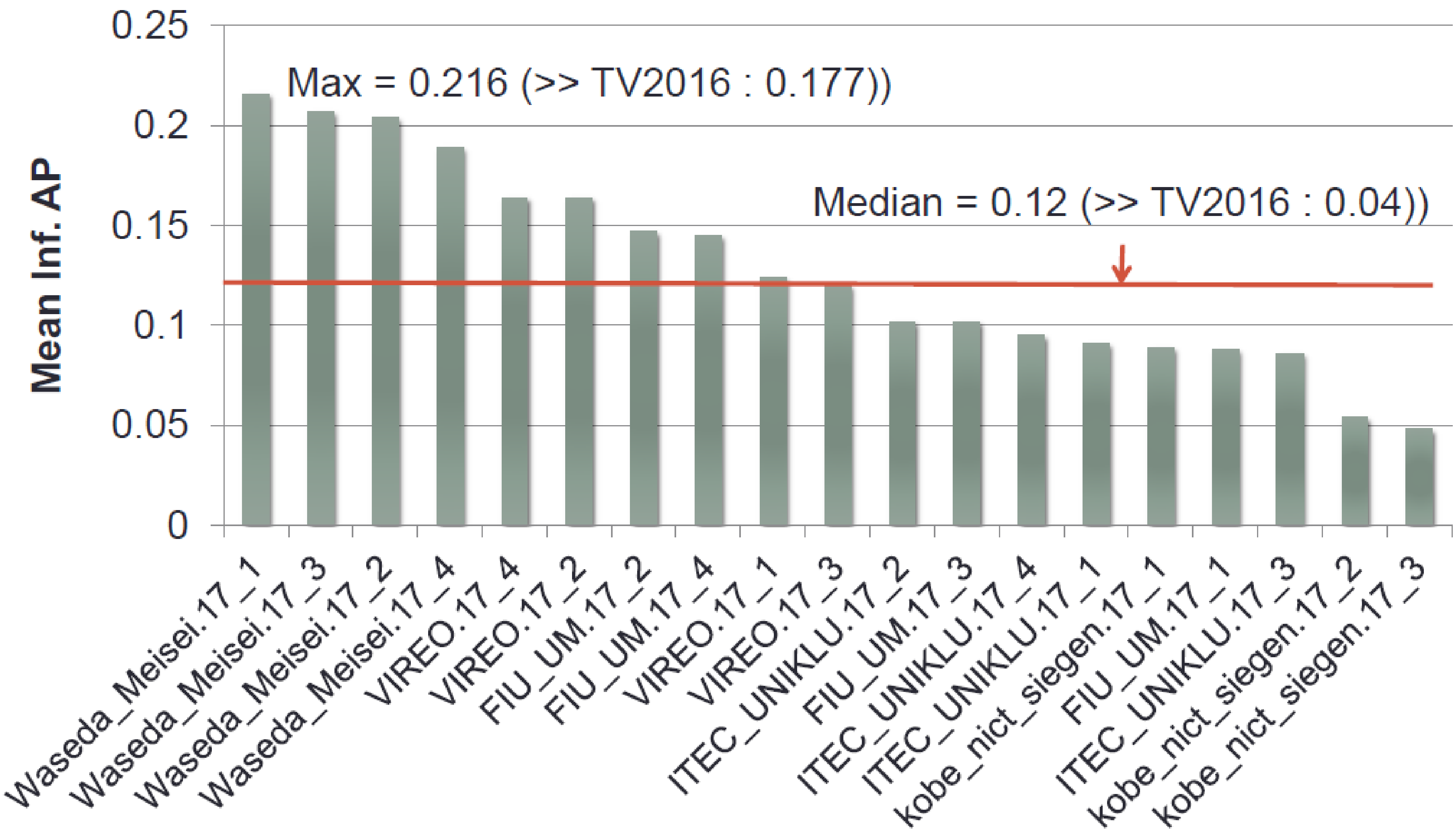
Each query assumed to be binary: absent or present for each master reference shot.

NIST sampled ranked pools and judged top results from all submissions.

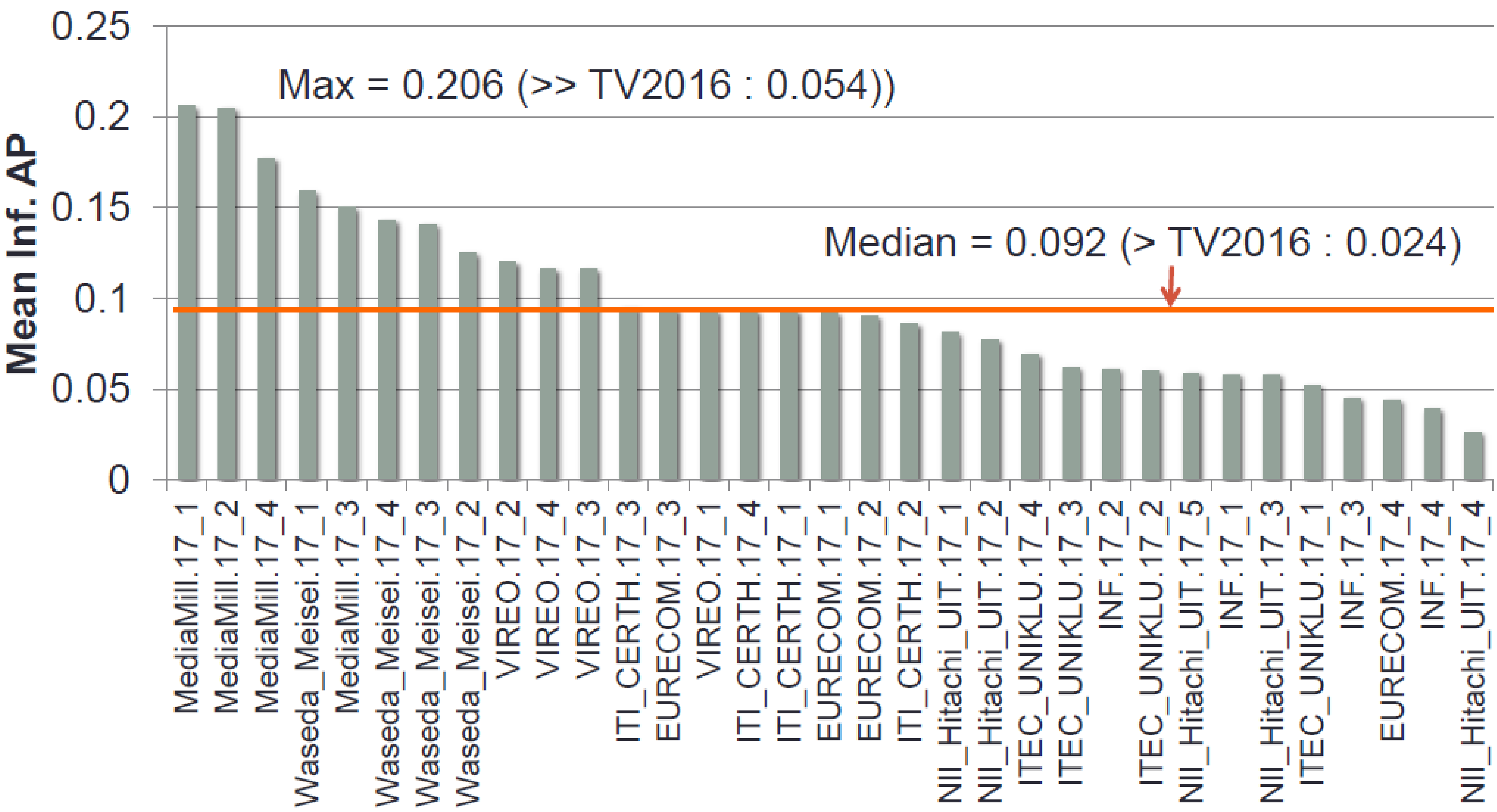
Metrics: *inferred average precision per query.*

Compared runs in terms of **mean** *inferred average precision* across the 30 queries.

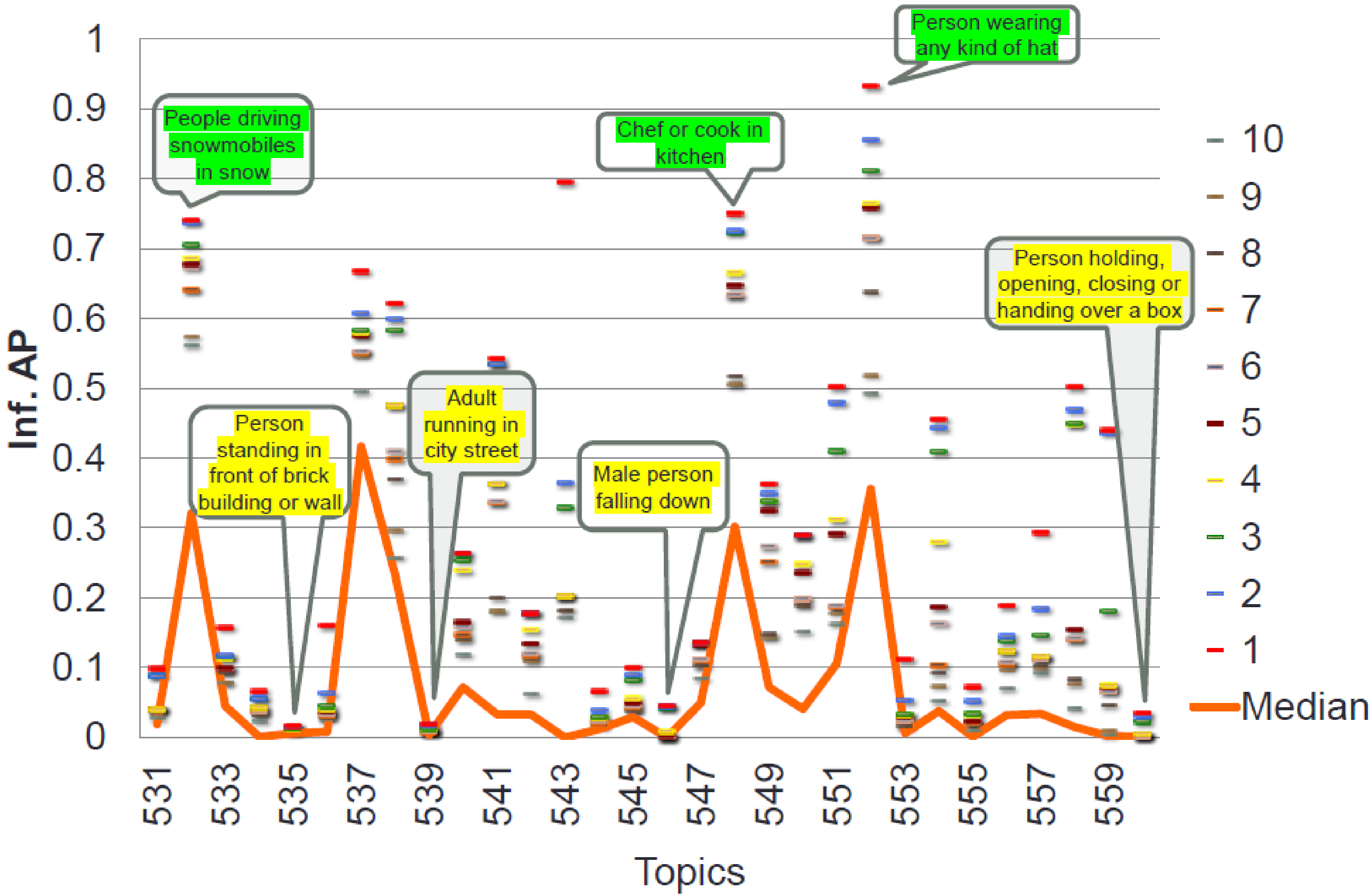
Submission scores for 19 manually assisted runs



Submission scores for 33 fully automatic runs



Top 10 infAP scores by query (fully automatic)



Which topics where easy or difficult overall?

Top 10 Easy (sorted by count of runs with InfAP >= 0.7)	Top 10 Hard (sorted by count of runs with InfAP < 0.7)
a person wearing any kind of hat	an adult person running in a city street
a chef or cook in a kitchen	person standing in front of a brick building or wall
one or more people driving snowmobiles in the snow	person holding, opening, closing or handing over a box
one or more people swimming in a swimming pool	a male person falling down
a man and woman inside a car	child or group of children dancing
a crowd of people attending a football game in a stadium	children playing in a playground
a newspaper	person talking on a cell phone
a person communicating using sign language	person holding or opening a briefcase
a person wearing a scarf	one or more people eating food at a table indoor
a person riding a horse including horse-drawn carts	person talking behind a podium wearing a suit outdoors during daytime

More action and dynamics in hard queries

Part II:
Results of submitted systems

Waseda_Meisei system

More than 50,000 concepts

[Step. 0] Preparation

Build a large semantic concept bank using pretrained convolutional neural networks (CNNs) and support vector machines (SVMs).

Our video retrieval pipeline consists of three steps:

[Step. 1]

Extract several search keywords based on the given query phrases. (manually or automatically)

[Step. 2]

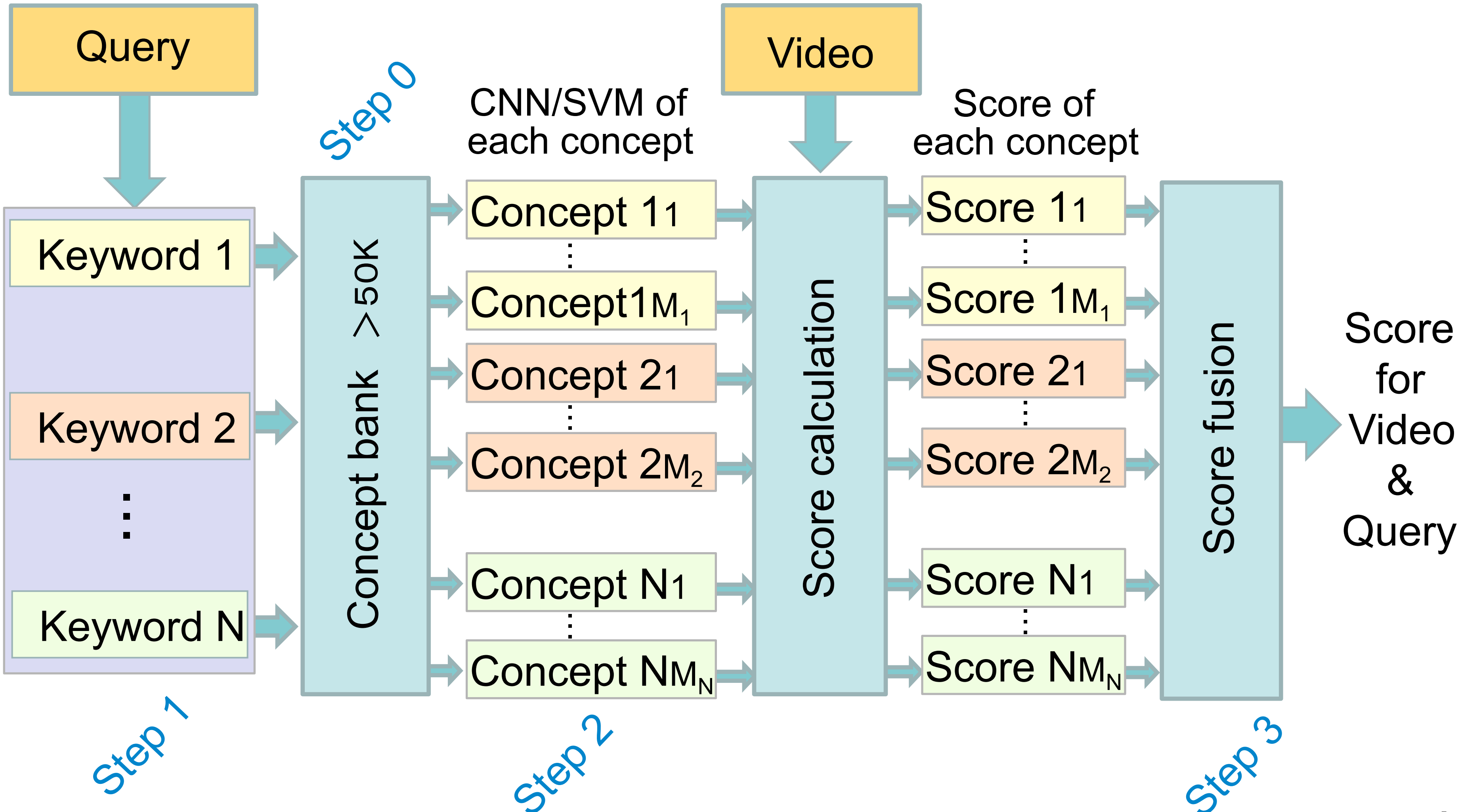
Choose concept classifiers based on selected keywords

[Step. 3]

Combine the semantic concept scores to obtain the final search result.

Waseda_Meisei system

"Find shots of one or more people driving snowmobiles in the snow"



Waseda_Meisei system [Step. 0]

Our concept bank for the TRECVID 2017 AVS task

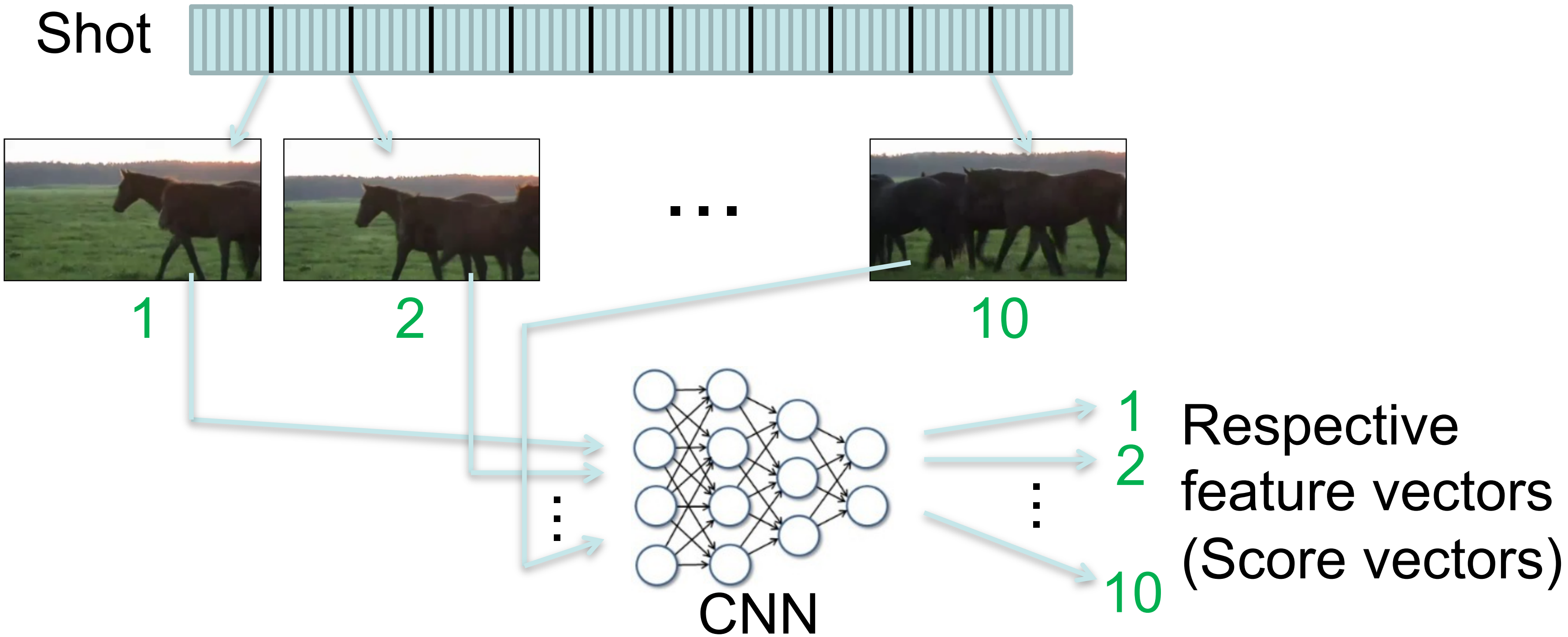
Name	Database	# of concepts	Concept type(s)
TRECVID346	TRECVID (ImageNet)	346	Object, Scene, Action
FCVID239	FCVID [4] (ImageNet)	239	Object, Scene, Action
UCF101	UCF101 [8] (ImageNet)	101	Action
PLACES205	Places [10]	205	Scene
PLACES365	Places	365	Scene
HYBRID1183	Places, ImageNet	1,183	Object, Scene
IMAGENET1000	ImageNet	1,000	Object
IMAGENET4000	ImageNet	4,000	Object
IMAGENET4437	ImageNet	4,437	Object
IMAGENET8201	ImageNet	8,201	Object
IMAGENET12988	ImageNet	12,988	Object
IMAGENET21841	ImageNet	21,841	Object

To provide good coverage for the given query phrases, we built a large concept bank consisting of more than 50,000 concepts.

Waseda_Meisei system [Step. 1]

Feature extraction

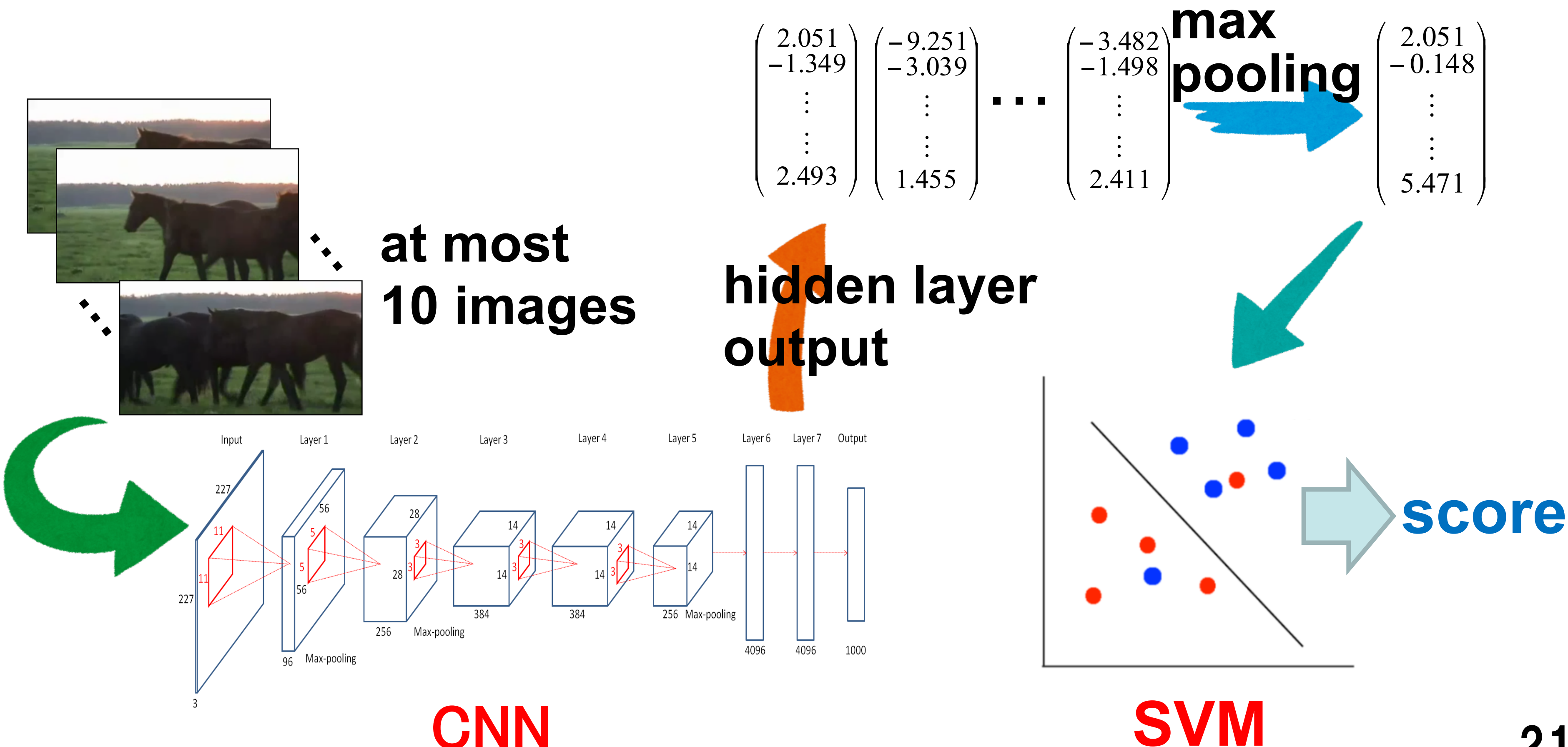
We selected at most 10 frames from each shot at regular intervals.



Waseda_Meisei system [Step. 1]

- TRECVID346
- FCVID239
- UCF101

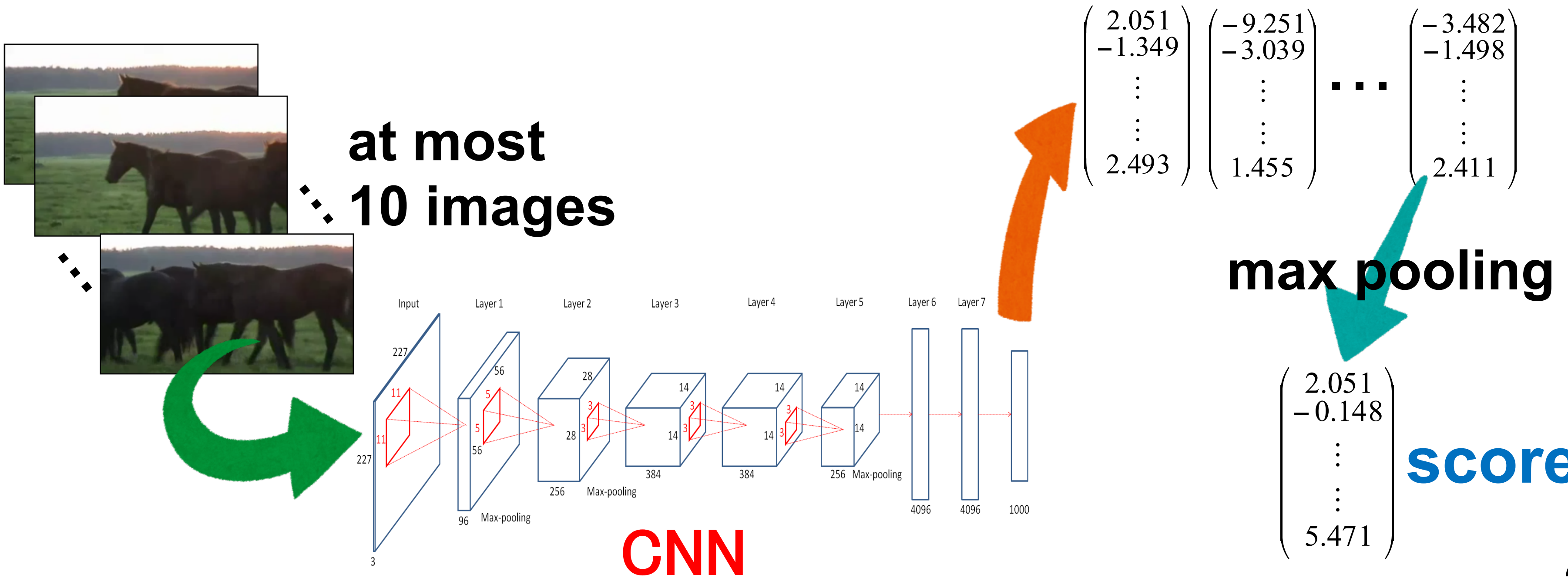
CNN/SVM tandem
connectionist architecture



Waseda_Meisei system [Step. 1]

PLACES205	IMAGENET1000	IMAGENET8201
PLACES365	IMAGENET4000	IMAGENET12988
HYBRID1183	IMAGENET4437	IMAGENET21841

The shot scores were obtained directly from the output layer (before softmax was applied)



Score normalization

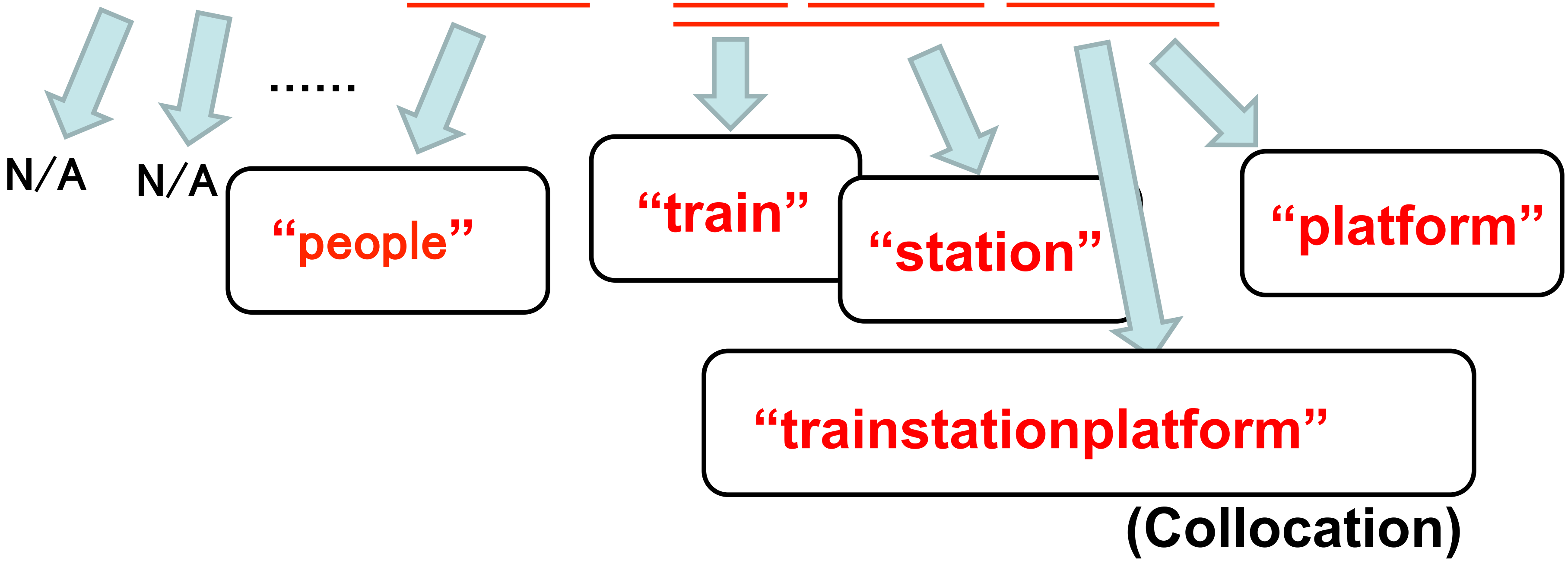
The score for each semantic concept was normalized over all test shots using a min-max normalization.

The maximum scores: 1.0 (most probable)
The minimum scores: 0.0 (least probable)

Waseda_Meisei system [Step. 1]

Extract keywords from a query.

Query:
“One or more people at train station platform”



Waseda_Meisei system [Step. 2]

Choose concept classifiers based on selected keywords

Query

At least two planes
Both visible

Keyword i

e.g. "plane"

e.g. "airplane" "aircraft"

Concept bank

Index 1 : Model of Concept 1

Index 2 : Model of Concept 2

Index 3 : Model of Concept 3

⋮

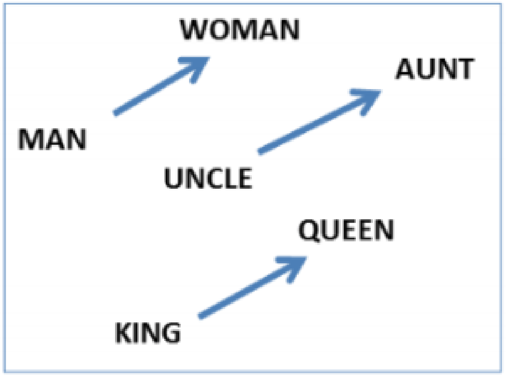
Index N : Model of Concept N

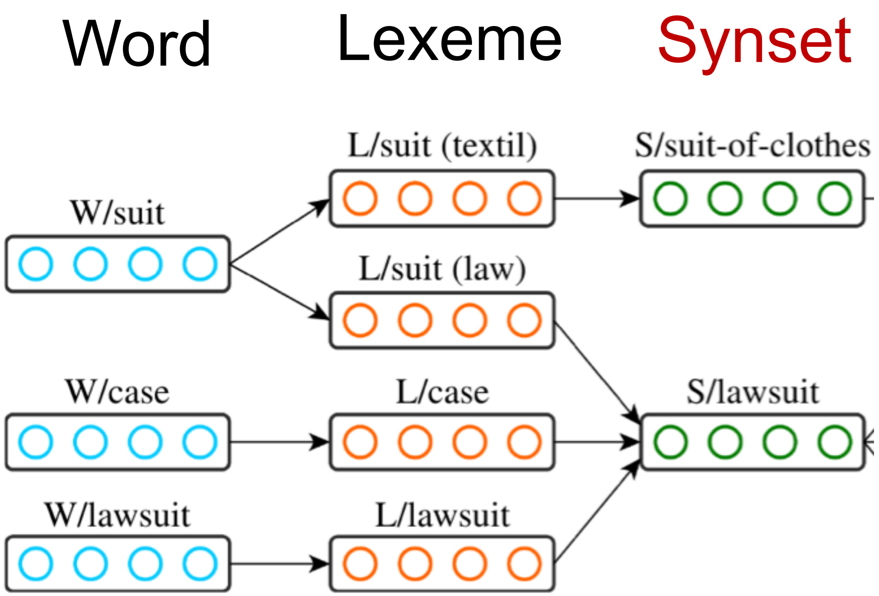
Problem:

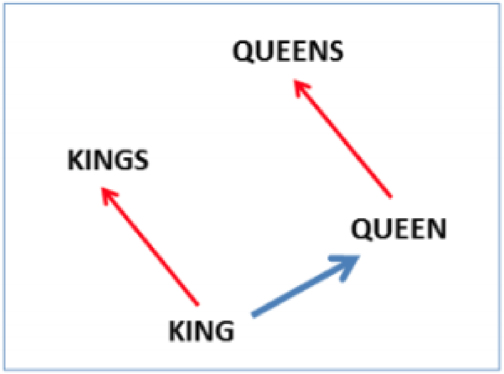
Representation of the keyword is not the same as that of the index word. Which concepts should be used for the keyword?

Waseda_Meisei system [Step. 2]

- WordNet based method
  - Exact match of *synset*.
- Word2Vec based method
  - Similarity of skip-gram.
- Hybrid of WordNet & Word2Vec







Waseda_Meisei system [Step. 2]

To deal with no-classifier concepts:

Semantically similar concept was chosen using the word2vec algorithm.

"phone" →

- **telephone**
- **cellulartelephone**
- **deskphone**
- **…**

Usually use a concept having cosine similarity ≥ 0.7

(depend on submitted runs)

Score fusion

Calculate the final scores by score-level fusion

	Multiply	Sum
w/o weight (*)	$\text{final scores} = \prod_{i=1}^{N} s_i$	$\text{final scores} = \sum_{i=1}^{N} s_i$
w/ weight	$\text{final scores} = \prod_{i=1}^{N} s_i^{w_i}$	$\text{final scores} = \sum_{i=1}^{N} w_i \cdot s_i$

(*) We used the IDF values calculated from the Microsoft COCO database as the fusion weights.

Waseda_Meisei system [Step. 3]

Multiply & w/o weight

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

$$\textbf{final scores} = \prod_{i=1}^{N} s_i$$

selected concepts

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.

Waseda_Meisei system [Step. 3]

Multiply & w/ weight

Almost the same as the previous method except for the incorporation of a fusion weight.

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

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: $(0.90)^{1.97} \times (0.70)^{8.23}$

$= 0.81 \times 0.05 = 0.04$

shot B: $(0.70)^{1.97} \times (0.90)^{8.23}$

$= 0.50 \times 0.42 = 0.21$

Waseda_Meisei system [Step. 3]

Sum & w/o weight

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

$$\text{final scores} = \sum_{i=1}^{N} s_i$$

“fountain” and “outdoor”

shot A:	0.70	+	0.10	=	0.80
shot B:	0.40	+	0.30	=	0.70
	⋮		⋮		⋮

Somewhat looser conditions than multiplying.

Waseda_Meisei system [Step. 3]

Sum & w/ weight

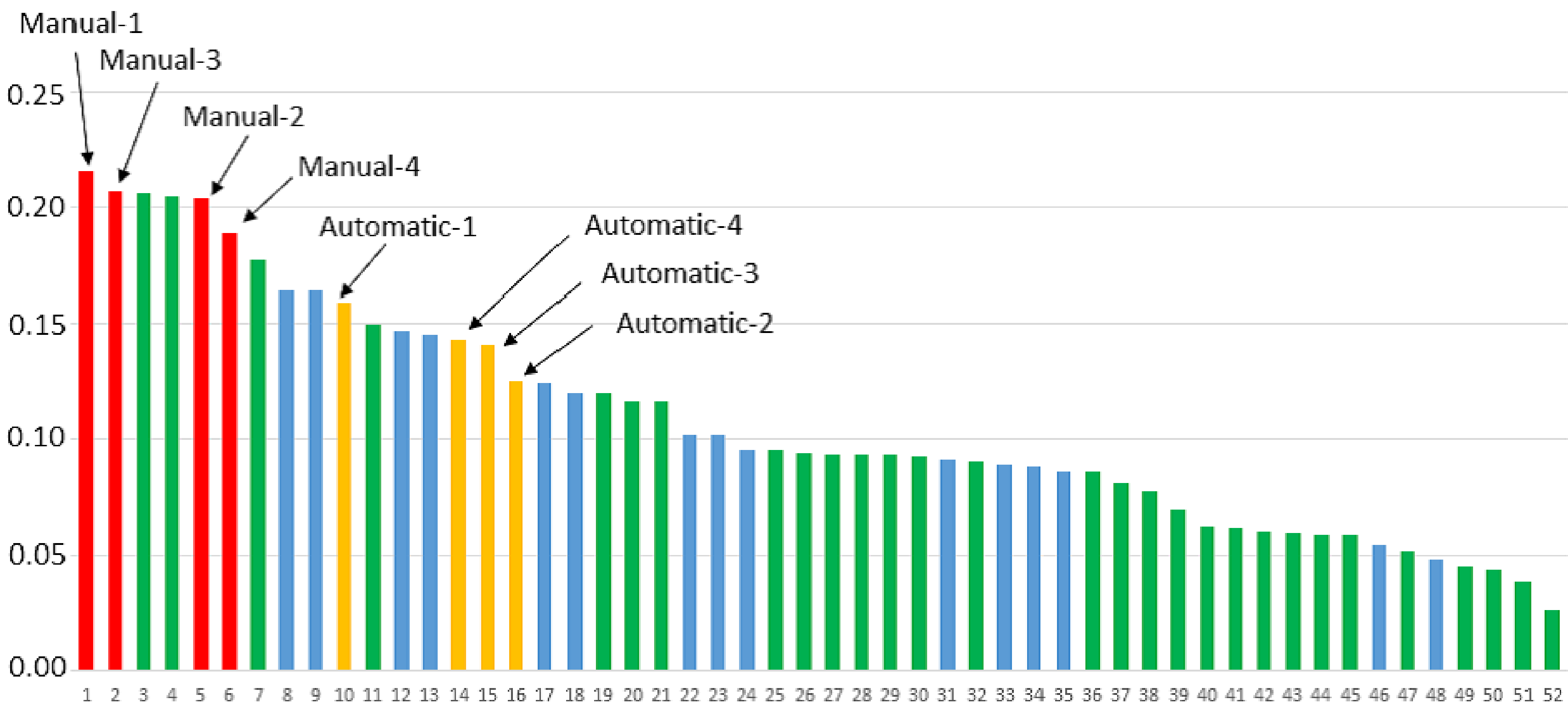
Summing weighted scores.

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

	"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

Waseda_Meisei system [Results]

Manual & Automatic runs

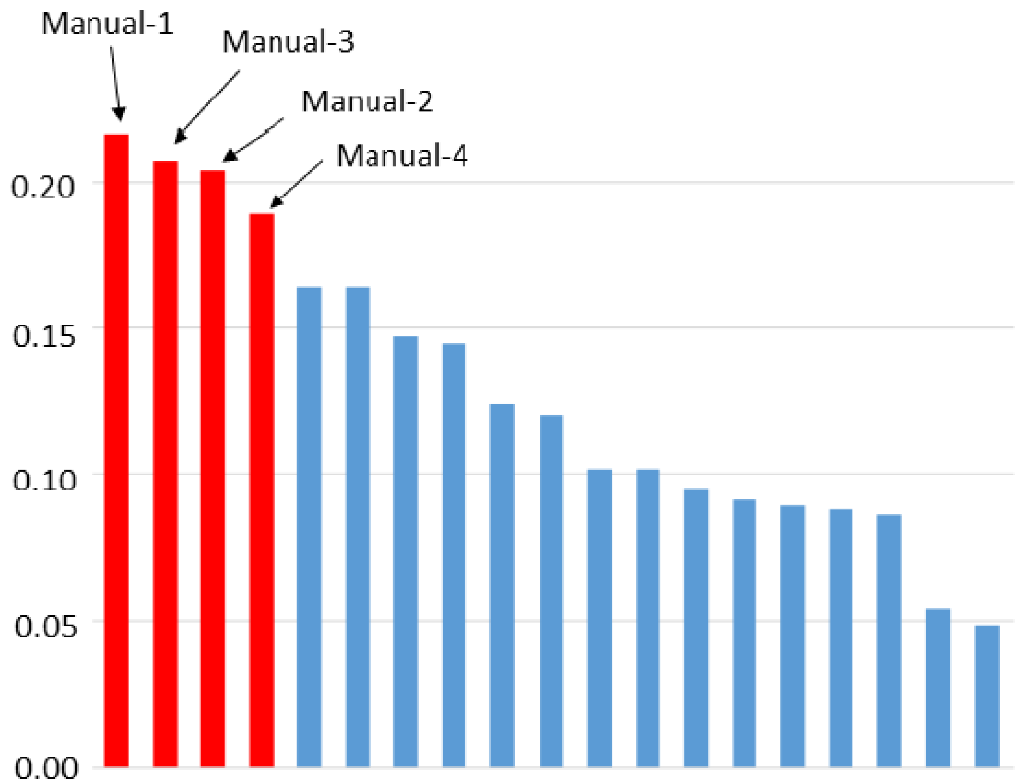


Comparison of WasedaMeisei runs with the runs of other teams
for all the submitted runs

Our Manual-1 run ranked 1st among the 52 runs.

Waseda_Meisei system [Results]

Manual runs

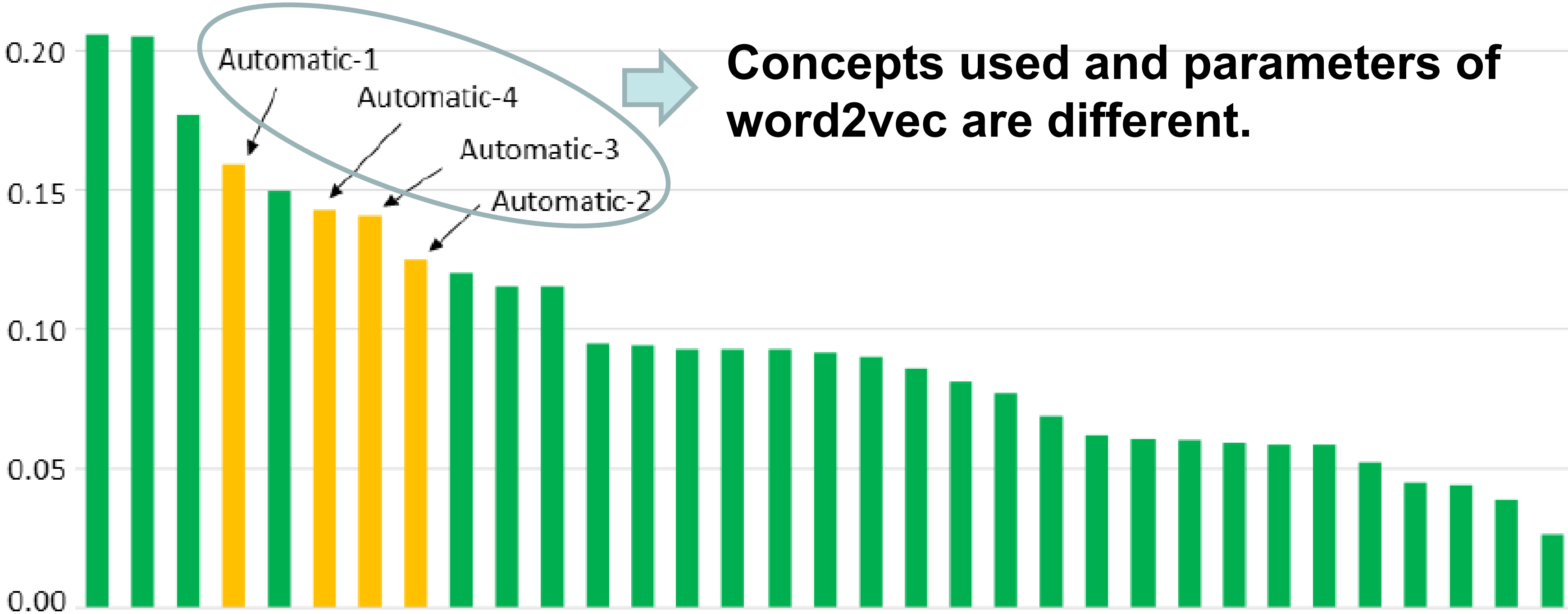


Comparison of Waseda Meisei runs with the runs of other teams for all submitted manually assisted runs.

Our manually assisted runs ranked 1st through the 4th overall.

Waseda_Meisei system [Results]

Automatic runs



Concepts used and parameters of word2vec are different.

Comparison of Waseda Meisei runs with the runs of other teams for all the fully automatic runs.

Our fully automatic runs ranked us 2nd overall among all participants.

Waseda_Meisei system [Results]

Comparison of Waseda_Meisei runs

Name	Fusion method	Fusion weight	mAP
Manual-1	Multiply	✓	21.6
Manual-2	Multiply		20.4
Manual-3	Sum	✓	20.7
Manual-4	Sum		18.9
Automatic-1	Multiply	✓	15.9
Automatic-2	Multiply	✓	12.5
Automatic-3	Multiply	✓	14.1
Automatic-4	Multiply	✓	14.3

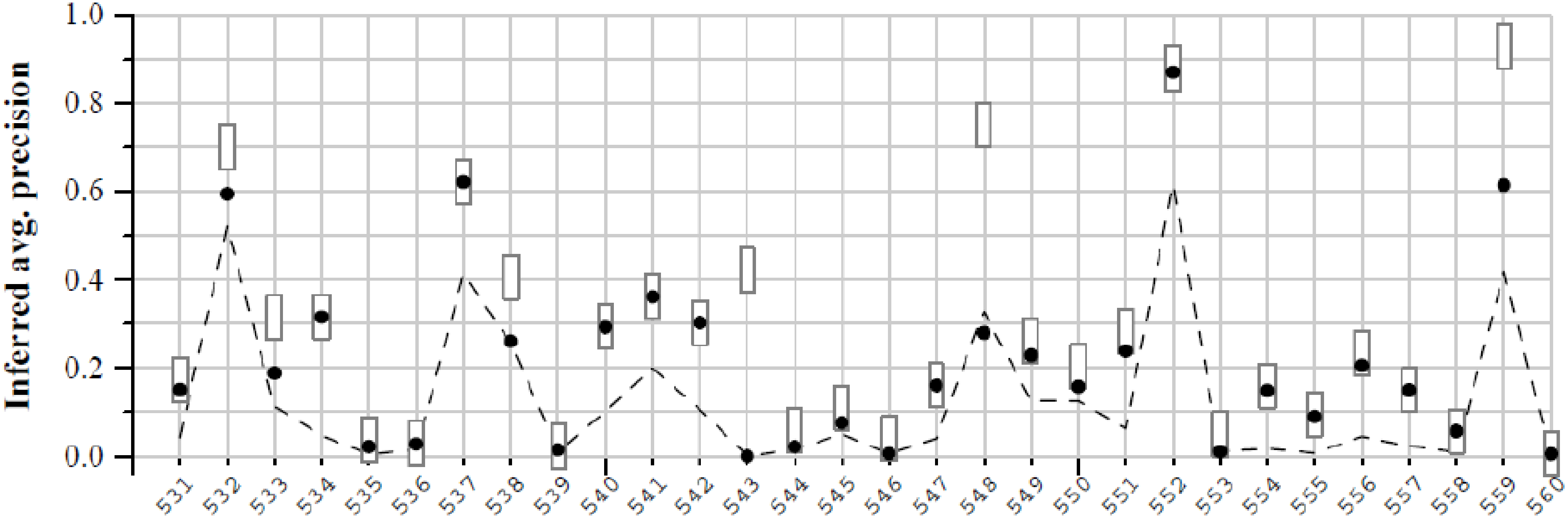
Manual vs. Automatic: Manual > Automatic

Fusion method: Multiply > Sum

Fusion weight: w/ weight > w/o weight

Waseda_Meisei system [Results]

Manual runs

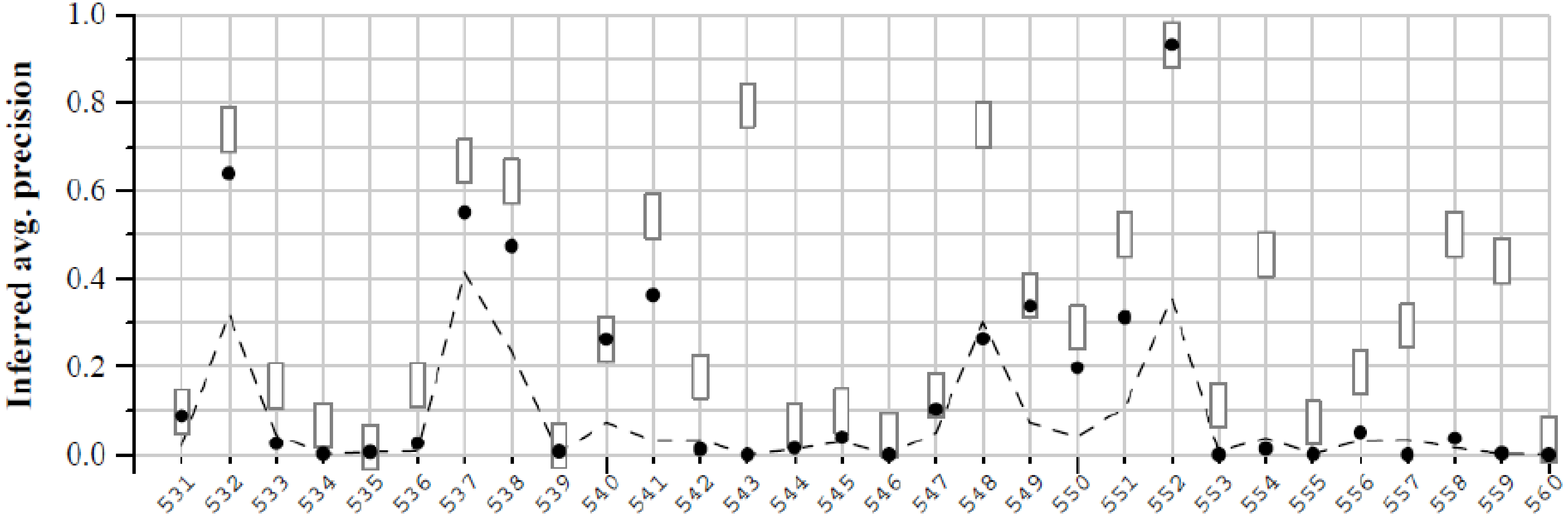


Average precision of our best manually assisted run (Manual1) for each query.
Run score (dot), median (dashes), and best (box) by query.

High performance was achieved by using a relatively large number of semantic concept classifiers (> 50,000).
The gap between the high and low performance widened;
average precisions for several query phrases were almost zero.

Waseda_Meisei system [Results]

Automatic runs



Average precision of our best fully automatic run (Automatic1) for each query.
Run score (dot), median (dashes), and best (box) by query.

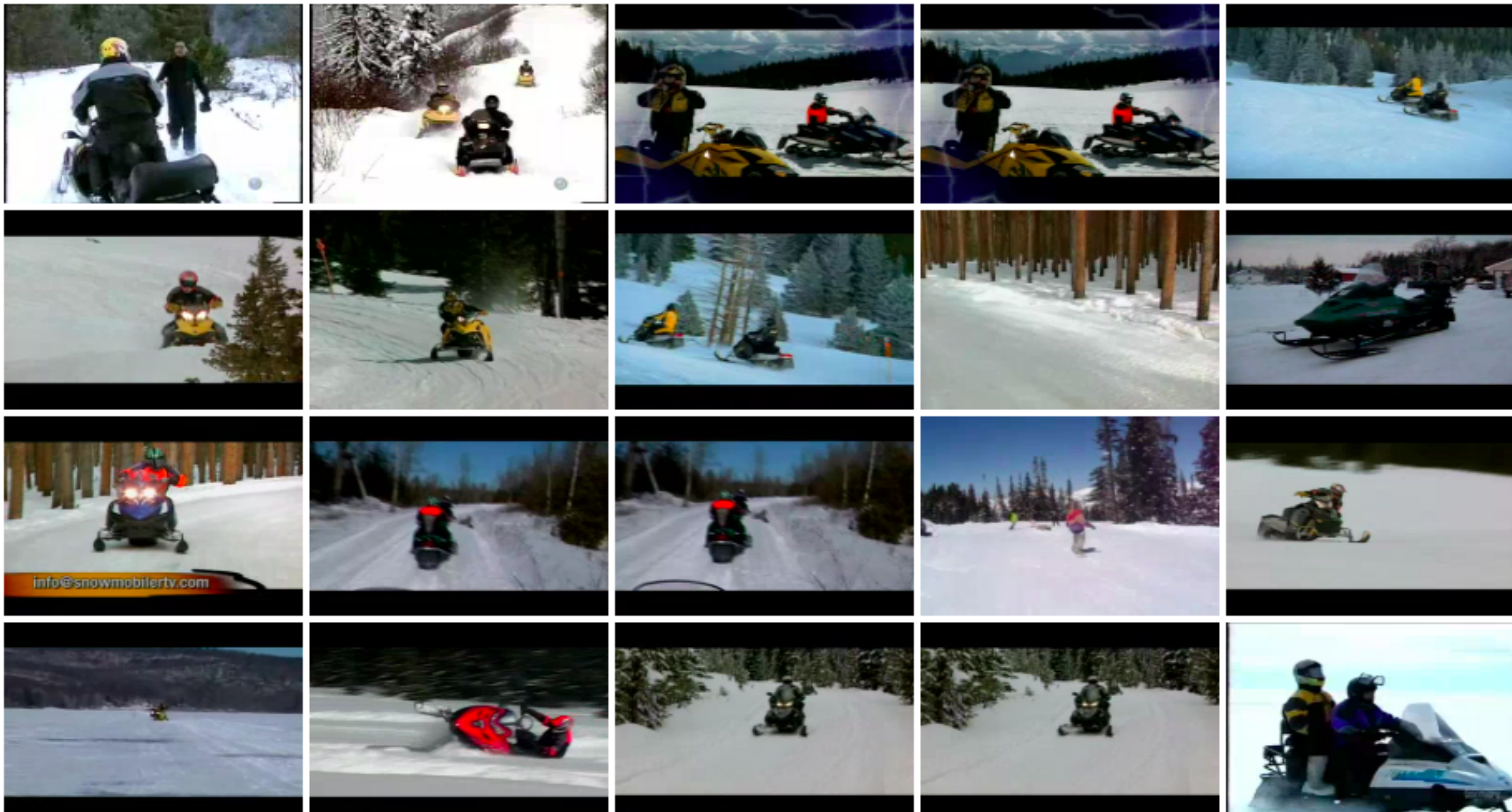
High performance was achieved by using a relatively large number of semantic concept classifiers (> 50,000).
The gap between the high and low performance widened;
average precisions for several query phrases were almost zero.

Waseda_Meisei system [Results]

Retrieved videos (manually-assisted system)

Good!!

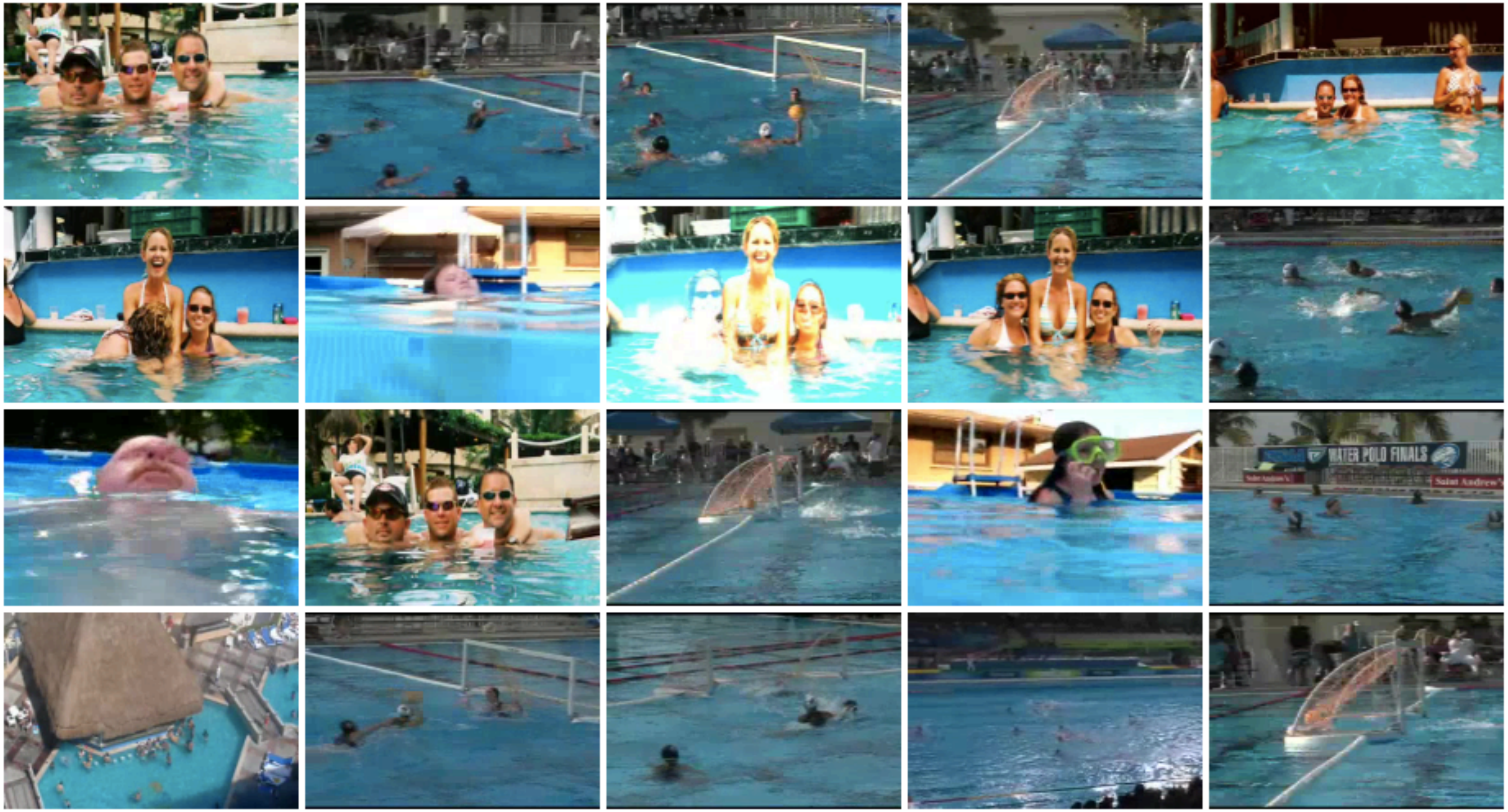
“one or more people driving snow mobiles in the snow”



Waseda_Meisei system [Results]

Retrieved videos (manually-assisted system) Good!!

"one or more people swimming in a swimming pool"



Waseda_Meisei system [Results]

Retrieved videos (fully-automatic system)

Bad…

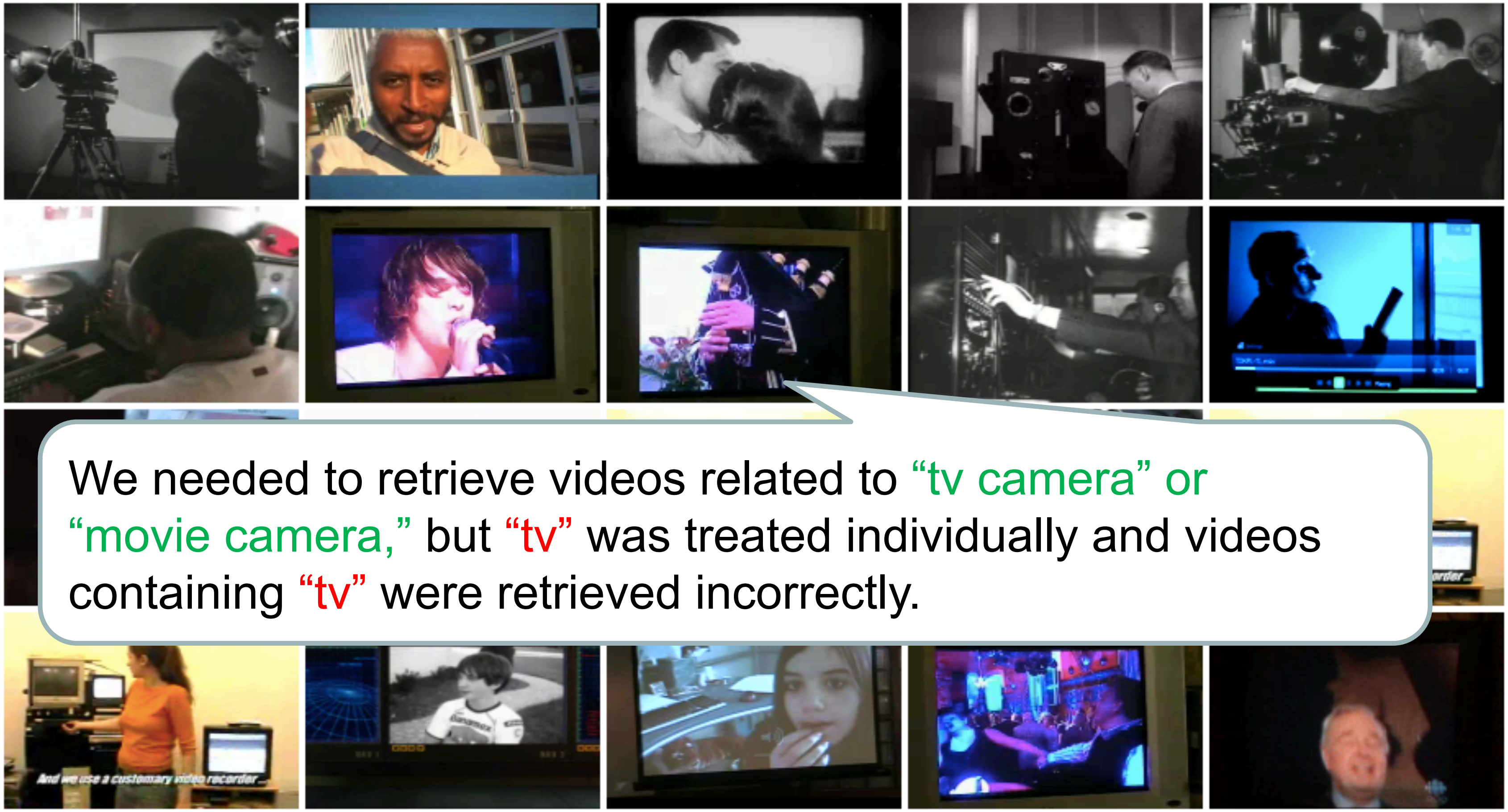
"a person holding or operating <u>a tv or movie camera</u>"

Waseda_Meisei system [Results]

Retrieved videos (fully-automatic system)

Bad…

"a person holding or operating <u>a tv or movie camera</u>"



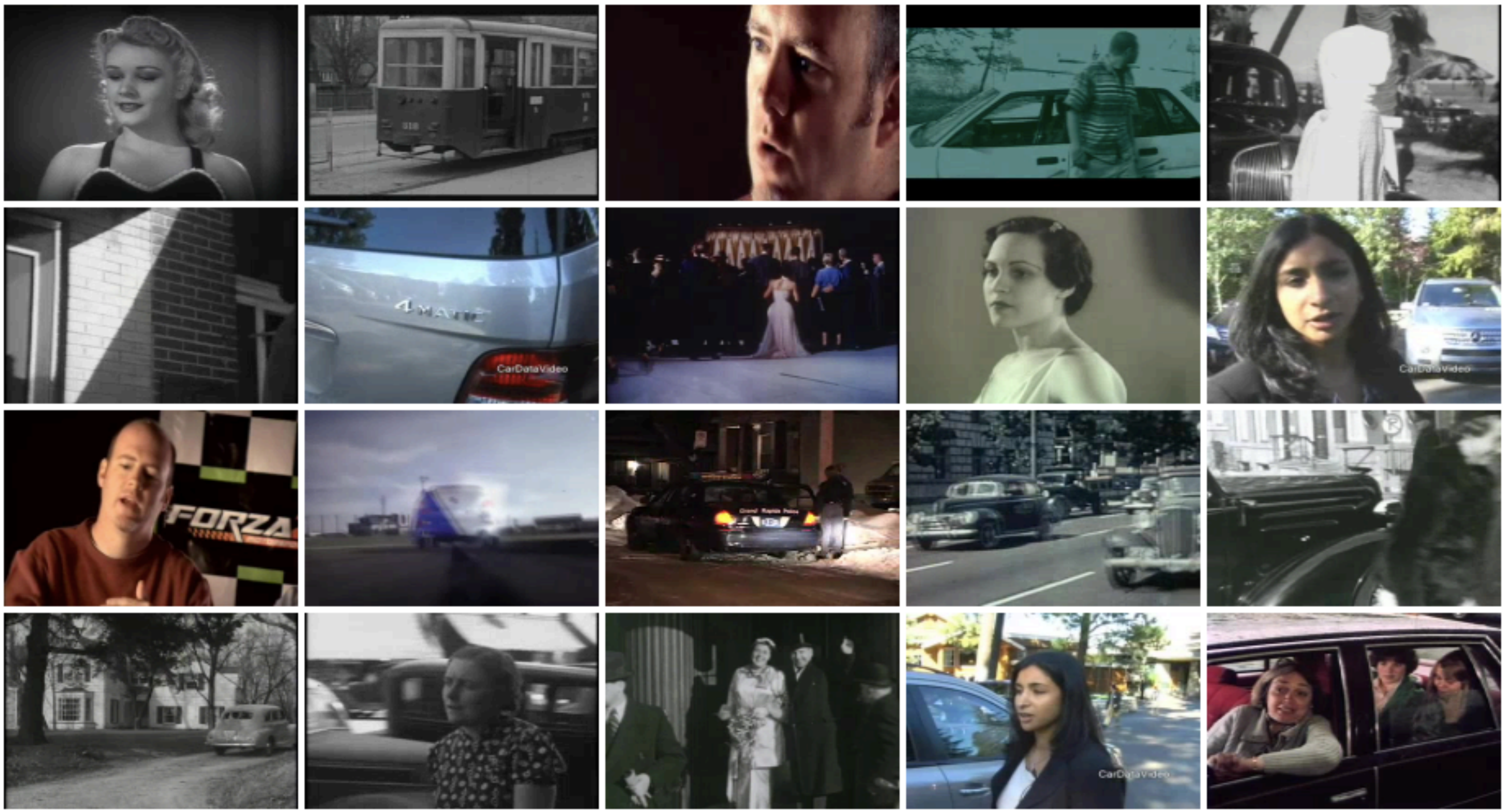
We needed to retrieve videos related to "tv camera" or "movie camera," but "tv" was treated individually and videos containing "tv" were retrieved incorrectly.

Waseda_Meisei system **[Results]**

Retrieved videos (fully-automatic system)

Bad…

“ a man and woman <u>inside a car</u>”



Waseda_Meisei system [Results]

Retrieved videos (fully-automatic system)

Bad…

“ a man and woman <u>inside a car</u>”

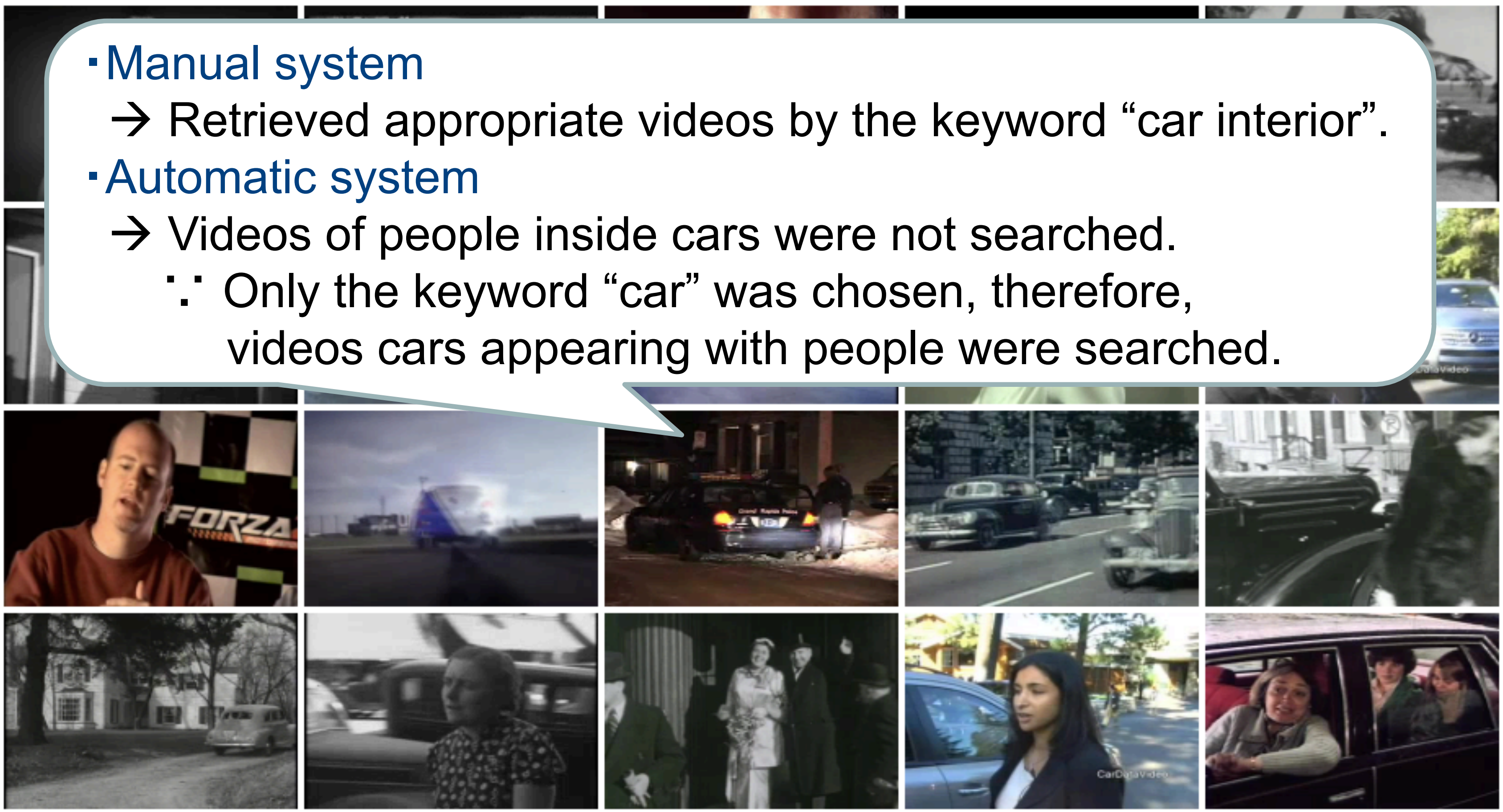
･Manual system

→ Retrieved appropriate videos by the keyword “car interior”.

･Automatic system

→ Videos of people inside cars were not searched.

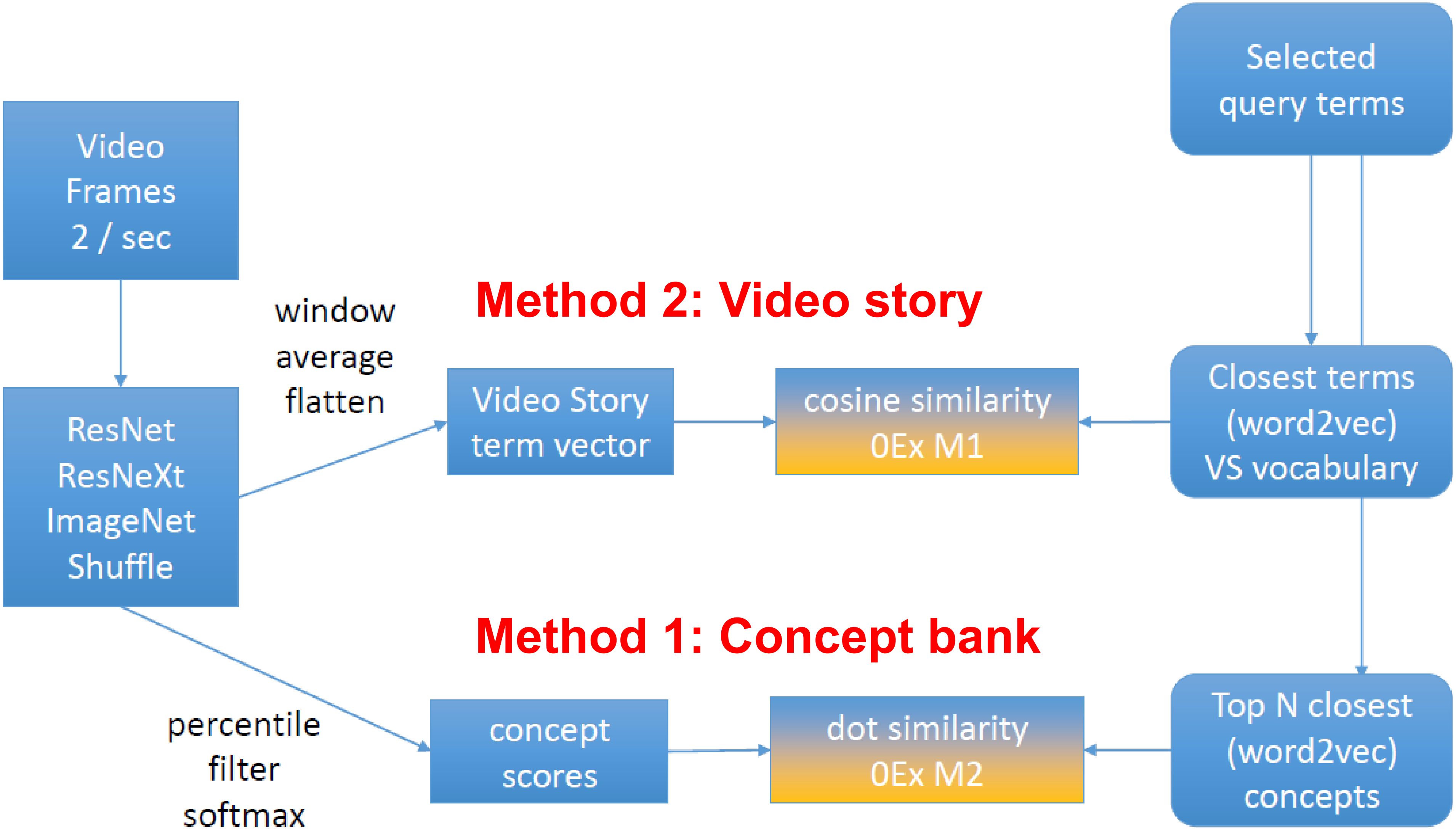
∵ Only the keyword “car” was chosen, therefore,
videos cars appearing with people were searched.



Waseda_Meisei system [Summary]

- We solved the problem of ad-hoc video search using a combination of many semantic concepts and selecting appropriate concepts from a concept bank that includes a wide variety of concepts.
- We achieved the best performance among all the submissions in 2017.
- However, the performance was still extremely poor for some query phrases.

MediaMill system [Pipeline]

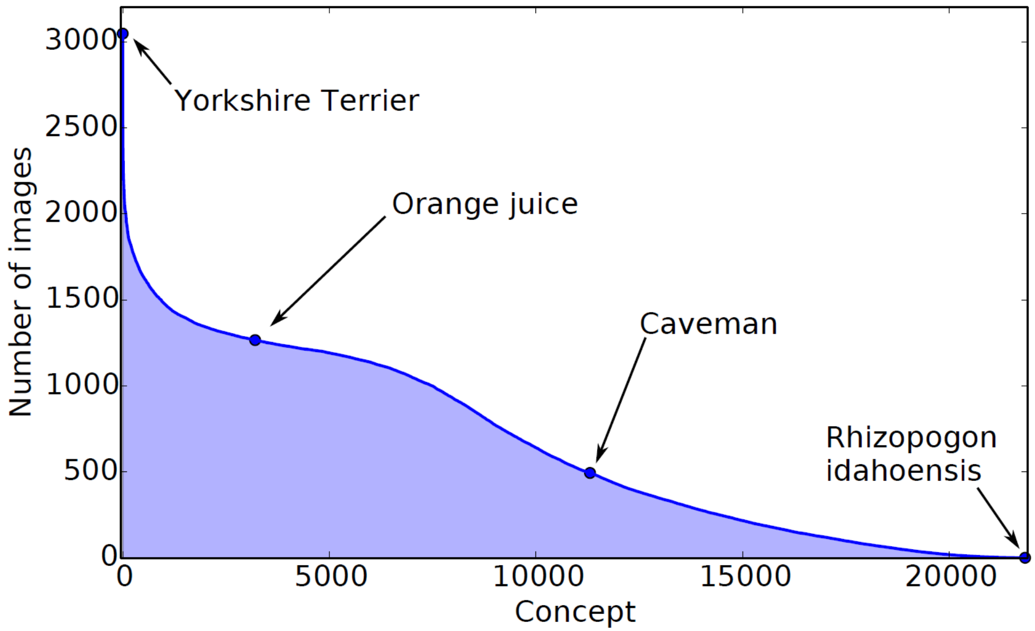


MediaMill system [Concept bank]

22k ImageNet classes

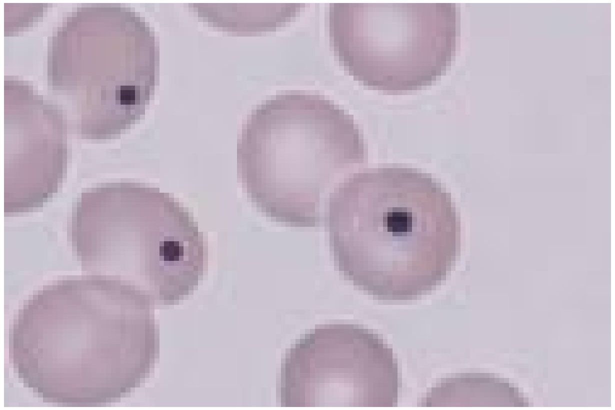
- Use as many classes as possible
- Find a balance between level of abstraction of classes and number of images in a class

Example imbalance

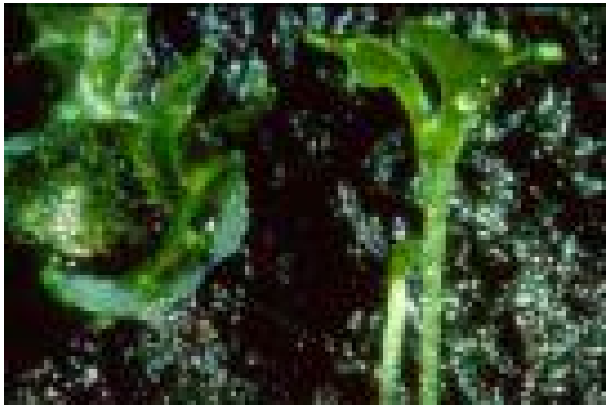


296 classes with 1 image

Irrelevant classes



Siderocyte

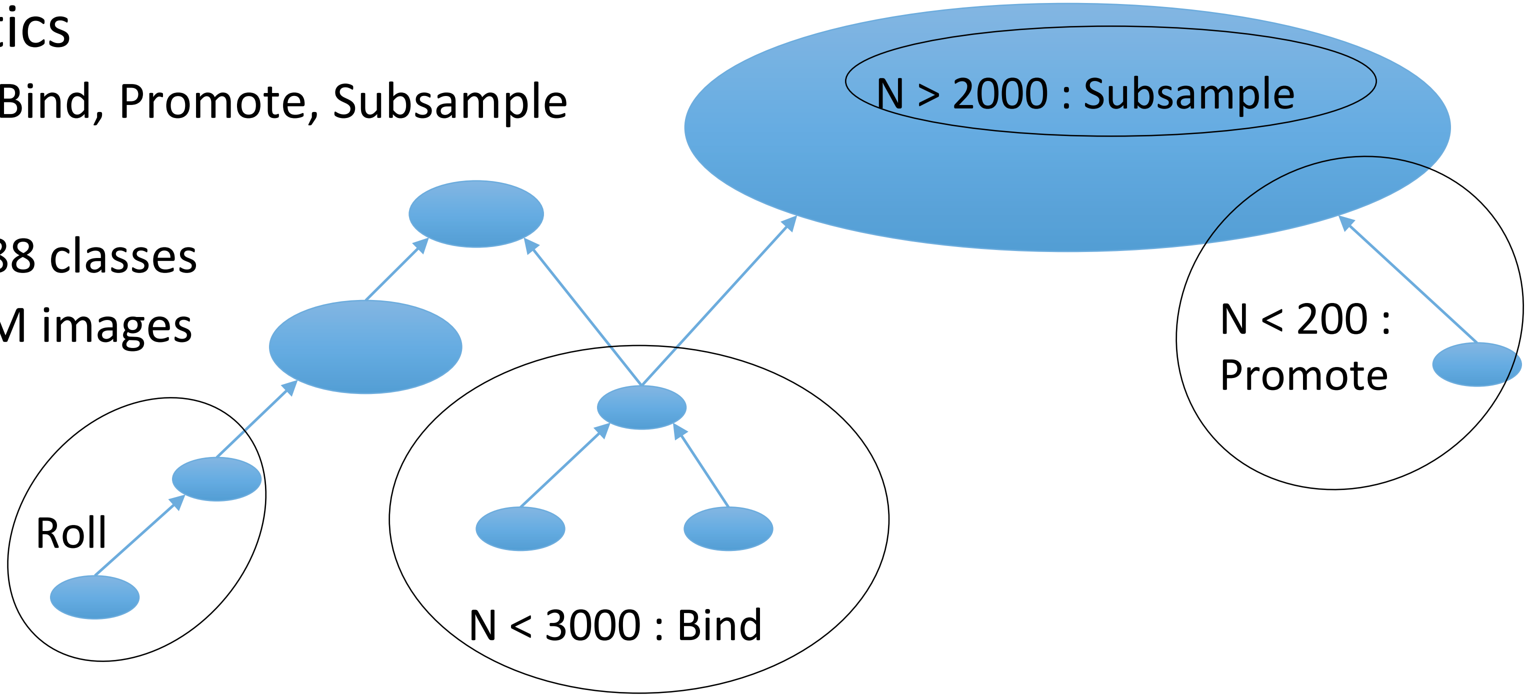


Gametophyte

MediaMill system [Concept bank]

CNN training on selection out of 22k ImageNet classes

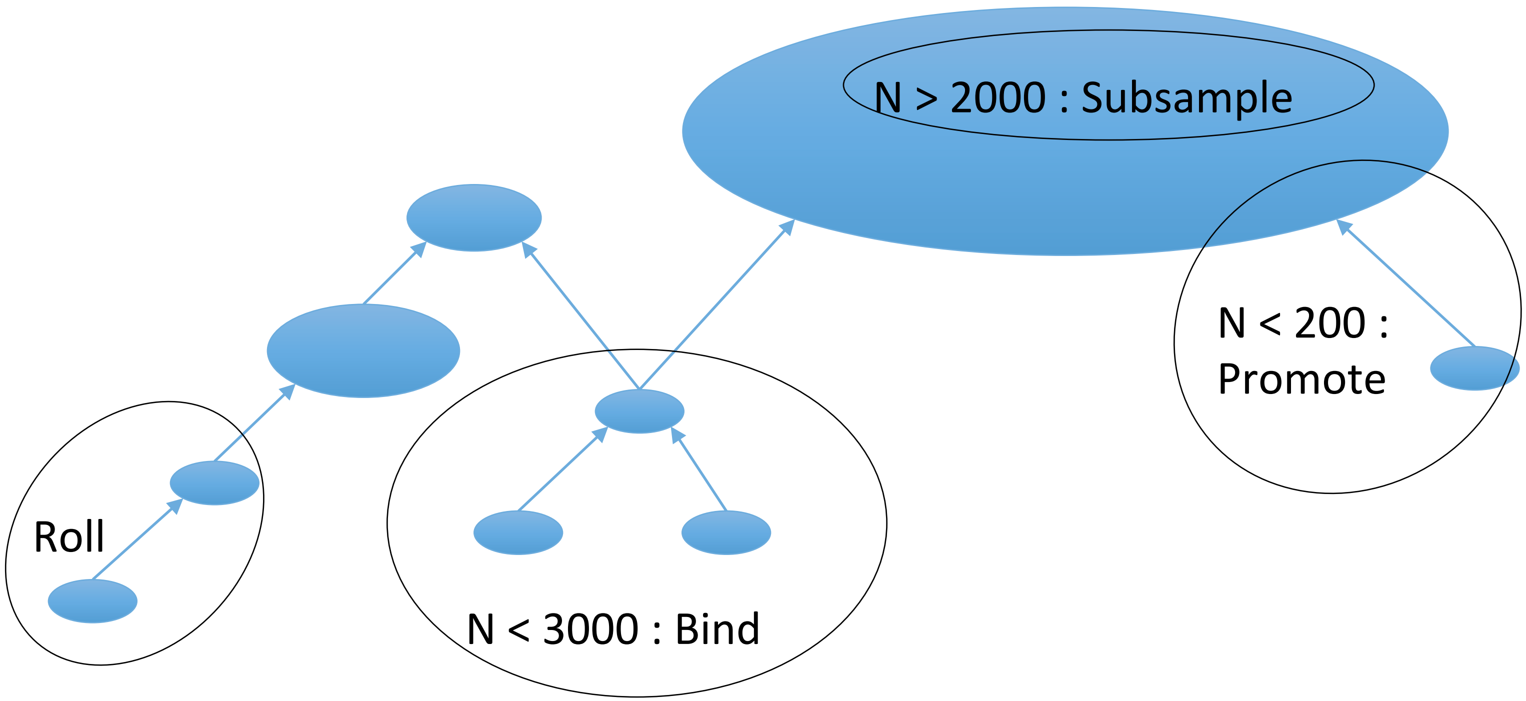
- Idea
  - Increase level of abstraction of classes
  - Incorporate classes with less than 200 samples
- Heuristics
  - Roll, Bind, Promote, Subsample
- Result
  - 12,988 classes
  - 13.6M images



The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection,
Pascal Mettes and Dennis Koelma and Cees Snoek,
International Conference on Multimedia Retrieval, 2016

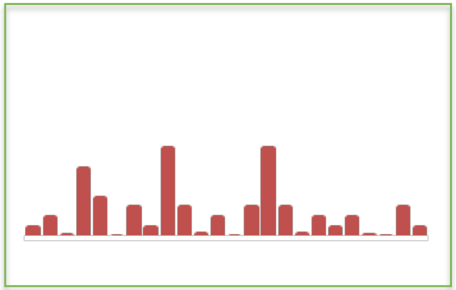
MediaMill system [Concept bank]

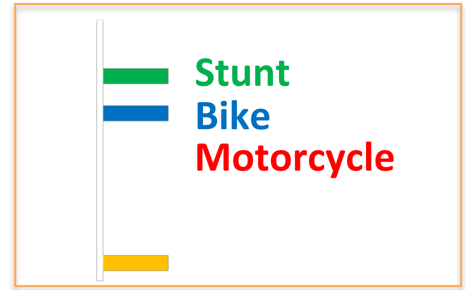
- Two networks
  - ResNet
  - ResNeXt
- Three datasets (subsets of ImageNet)
  - Roll Bind (3000) Promote (200) Subsample, 13k classes, training: 1000 images/class
  - Roll Bind (7000) Promote (1250) Subsample, 4k classes, training: 1706 images/class
  - Top 4000 classes, Breadth-first search >1200 images, training: 1324 images/class



MediaMill system [Video story]

Embed the story of a video





x_i s_i y_i

W — visual projection matrix → **Embedding** ← A — textual projection matrix

Joint optimization of W and A to preserve

Descriptiveness: preserve video descriptions : L(A,S)

Predictability: recognize terms from video content : L(S,W)

Videostory: A new multimedia embedding for few-example recognition and translation of events,
Amirhossein Habibian and Thomas Mensink and Cees Snoek,
Proceedings of the ACM International Conference on Multimedia, 2014

Video Story Training Sets

- VideoStory46k - www.mediamill.nl
  - 45826 videos from YouTube based on 2013 MED research set terms
- FCVID: Fudan Columbia Video Dataset
  - 87609 videos
- EventNet
  - 88542 videos
- Merged (VideoStory46k, FCVID, EventNet)

- Video Story dictionary: Terms that occur more than 10 times in the dataset
  - Merged : 6440 terms
- Using vocabulary of stemmed terms that occur more than 100 times in Wikipedia dump
  - With stemming: Respect the Video Story dictionary
  - 267.836 terms
- Use word2vec to expand them per video

MediaMill system

Query Terms

- Experiments show it is important to select the right terms
  - Instead of just taking the average of the terms in word2vec space
- Part-of-Speech tagging
  - <noun1> , <verb> , <noun2>
  - <subject> , <predicate> , <remainder>
- Query Plan

  A. Use nouns, verbs, and adjectives in <subject>
  - unless it concerns a person (noun1 = “person”, ”man”, “woman”, “child”, …)

  B. Use nouns in <remainder>
  - unless it concerns a person or noun is a setting (“indoors”, “outdoors”, …)

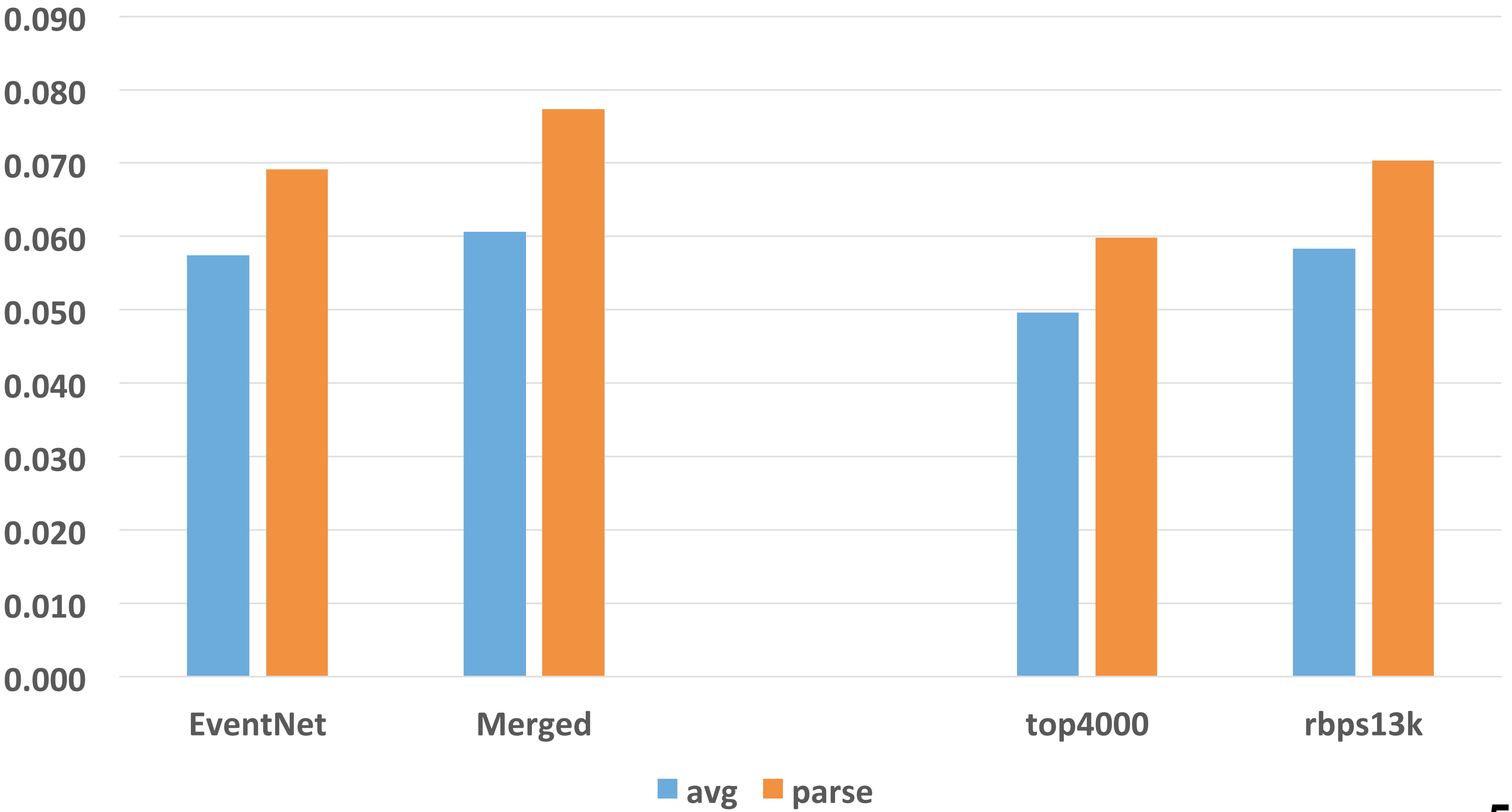
  C. Use <predicate>

  D. Use all nouns in sentence
  - Unless noun is a person or a setting

MediaMill system

The Effect of Parsing on 2016 Topics

- MIAP using only ResNet feature



MediaMill system

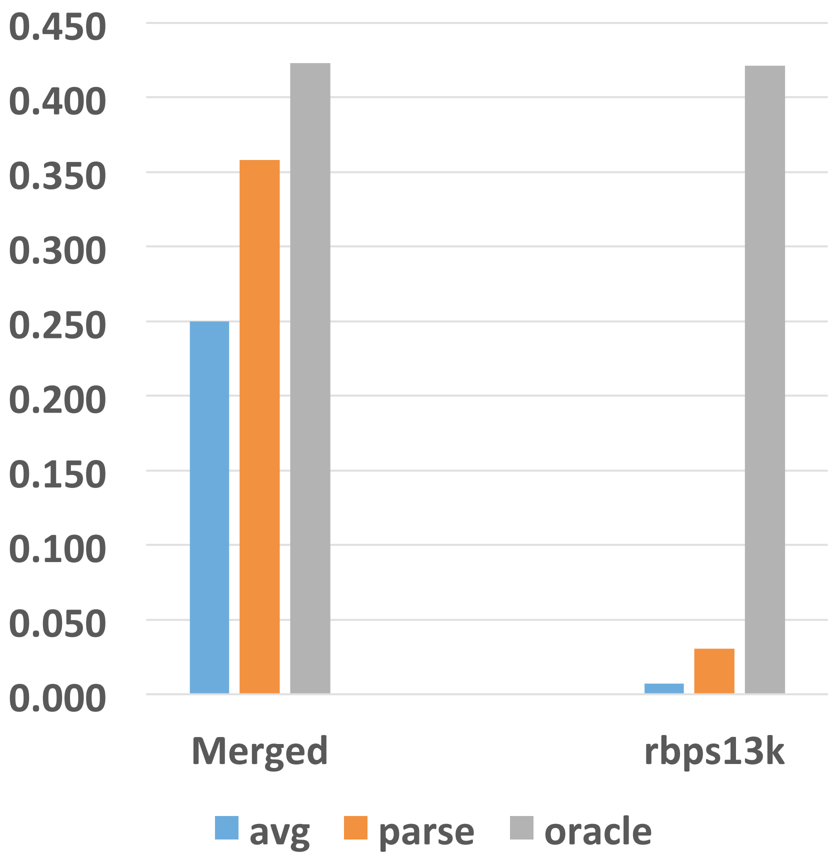
(Greedy) Oracle on 2016 Topics

- Fuse top (max 5) words/concepts with highest MIAP
- MIAP using only ResNet feature

MediaMill system

Query Examples : The Good

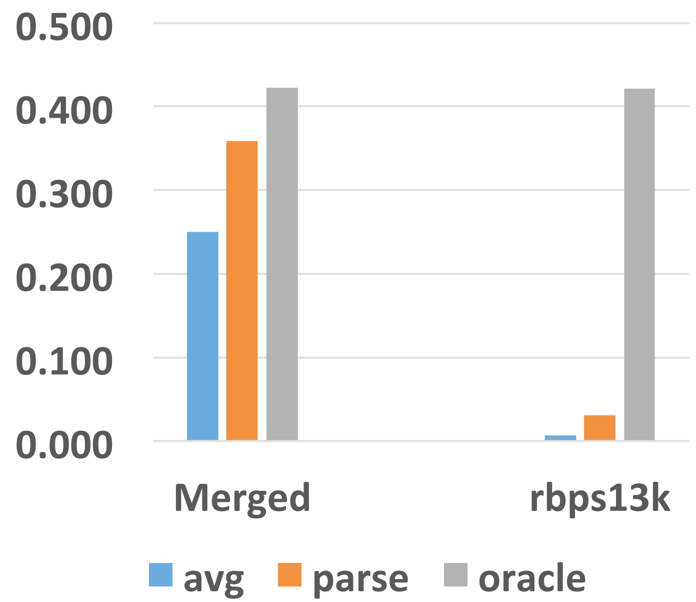
- A person playing drums indoors
- VideoStory terms avg :
  - person
  - plai
  - drum
  - indoor
- VideoStory terms parse :
  - drum
- VideoStory terms oracle :
  - beat
  - drum
  - snare
  - vibe
  - bng



MediaMill system

Query Examples : The Ambiguous

- A person playing drums indoors
- Concepts top5 avg :
  - guitarist, guitar player
  - outdoor game
  - drum, drumfish
  - sitar player
  - brake drum, drum
- Concepts top5 parse :
  - drum, drumfish
  - brake drum, drum
  - barrel, drum
  - snare drum, snare, side drum
  - drum, membranophone, tympan



Oracle :
- percussionist
- cymbal
- drummer
- drum, membranophone, tympan
- snare drum, snare, side drum

MediaMill system

Query Examples : The Bad

- A person sitting down with a laptop visible
- VideoStory terms avg :

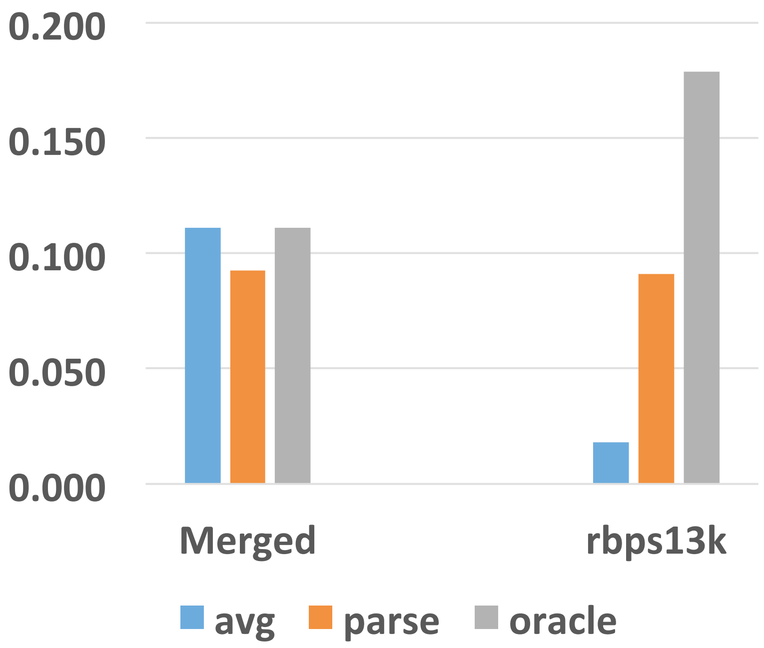
    person
    sit
    laptop

- VideoStory terms parse :

    laptop

- VideoStory terms oracle :

    monitor
    aspir
    acer
    alienwar
    vaio
    asus
    laptop (rank 7)



MediaMill system

Query Examples : The Difficult

- A person wearing a helmet
- Concept top5 parse :

  helmet (a protective headgear made of hard material to resist blows)

  helmet (armor plate that protects the head)

  pith hat, pith helmet, sun helmet, topee, topi

  batting helmet

  crash helmet

- Concept top5 oracle :

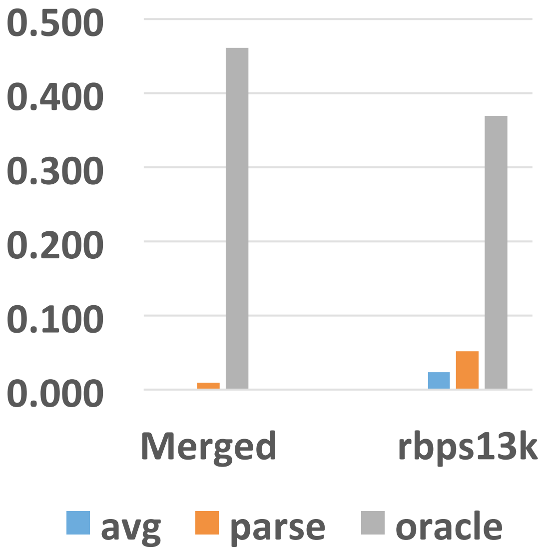
  hockey skate

  hockey stick

  ice hockey, hockey, hockey game

  field hockey, hockey

  rink, skating rink



Query Examples : The Impossible

- A crowd demonstrating in a city street at night
  - Parsing “fails”
  - Average wouldn’t have helped
- VS oracle :

  vega

  squar

  gang

  times

  occupi
- Concept oracle :

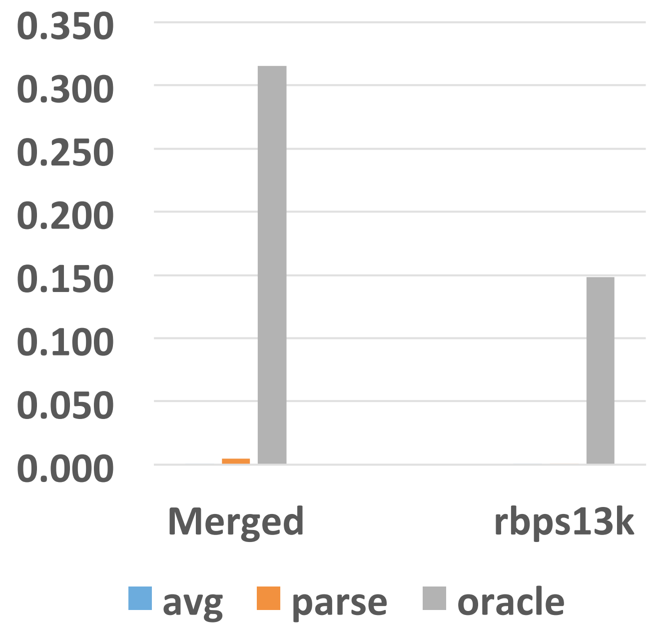
  vigil light, vigil candle

  motorcycle cop, motorcycle policeman, speed cop

  rider

  minibike, motorbike

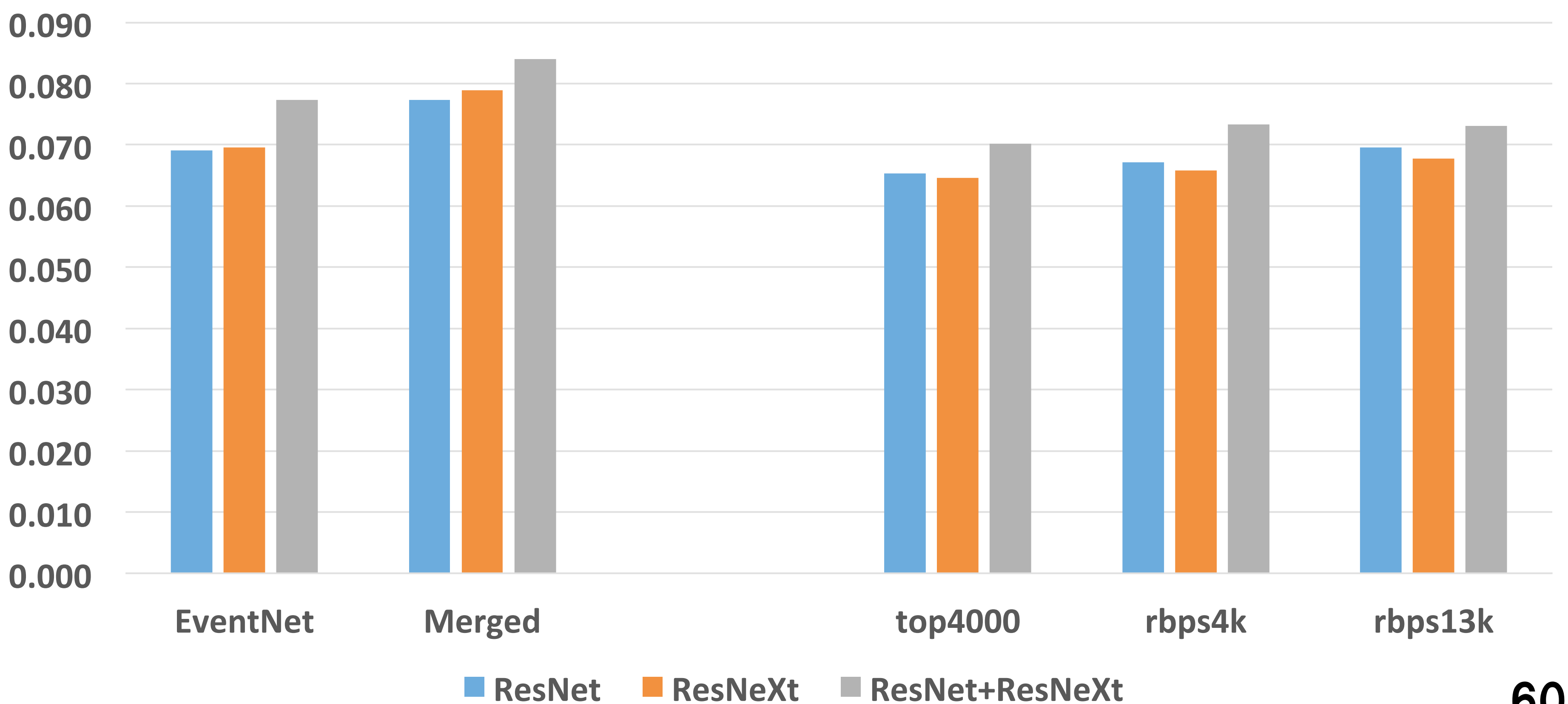
  freewheel



MediaMill system

Results 5 Modalities x 2 Features

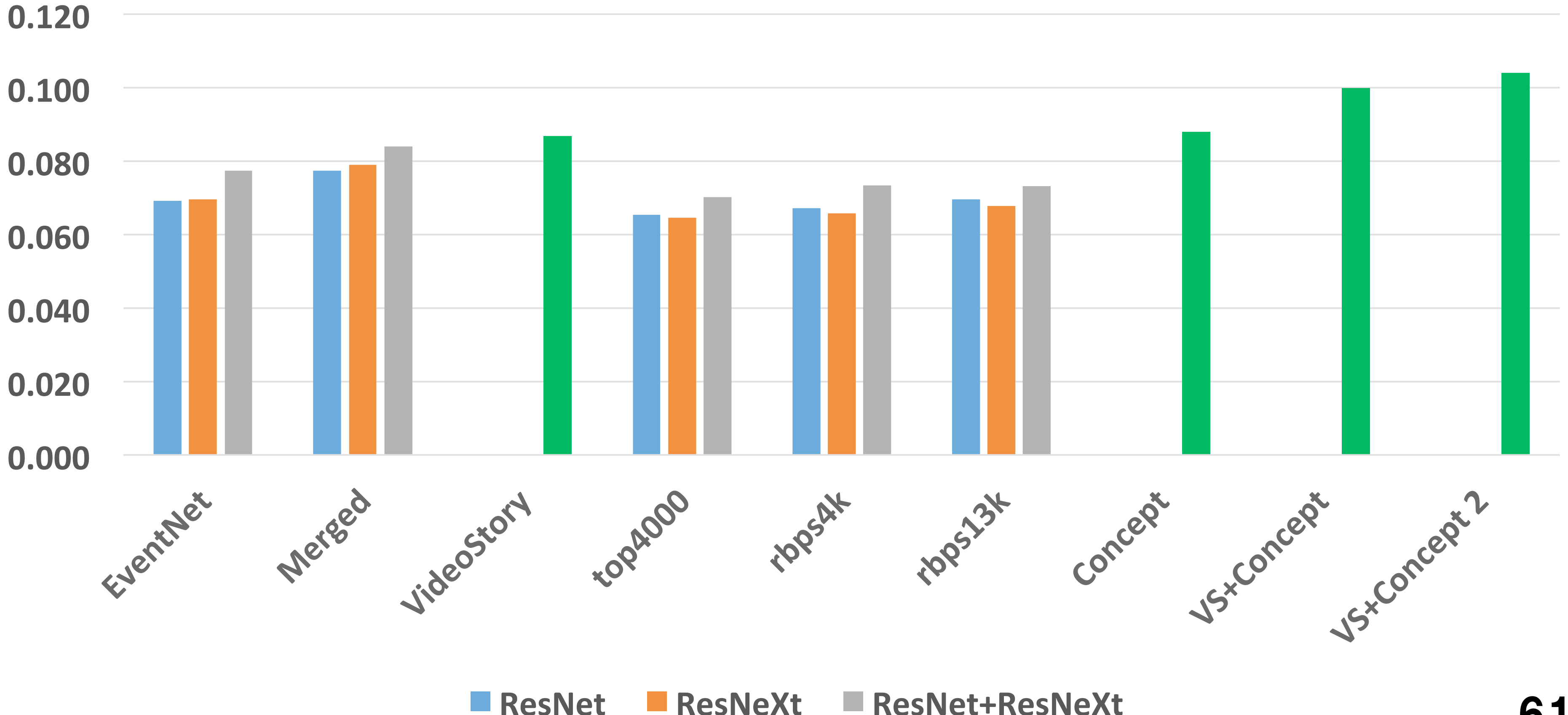
- VideoStory : ResNeXt is better than ResNet
- Concepts : ResNet is better than ResNeXt (overfit?)
- VideoStory is better than Concepts



MediaMill system

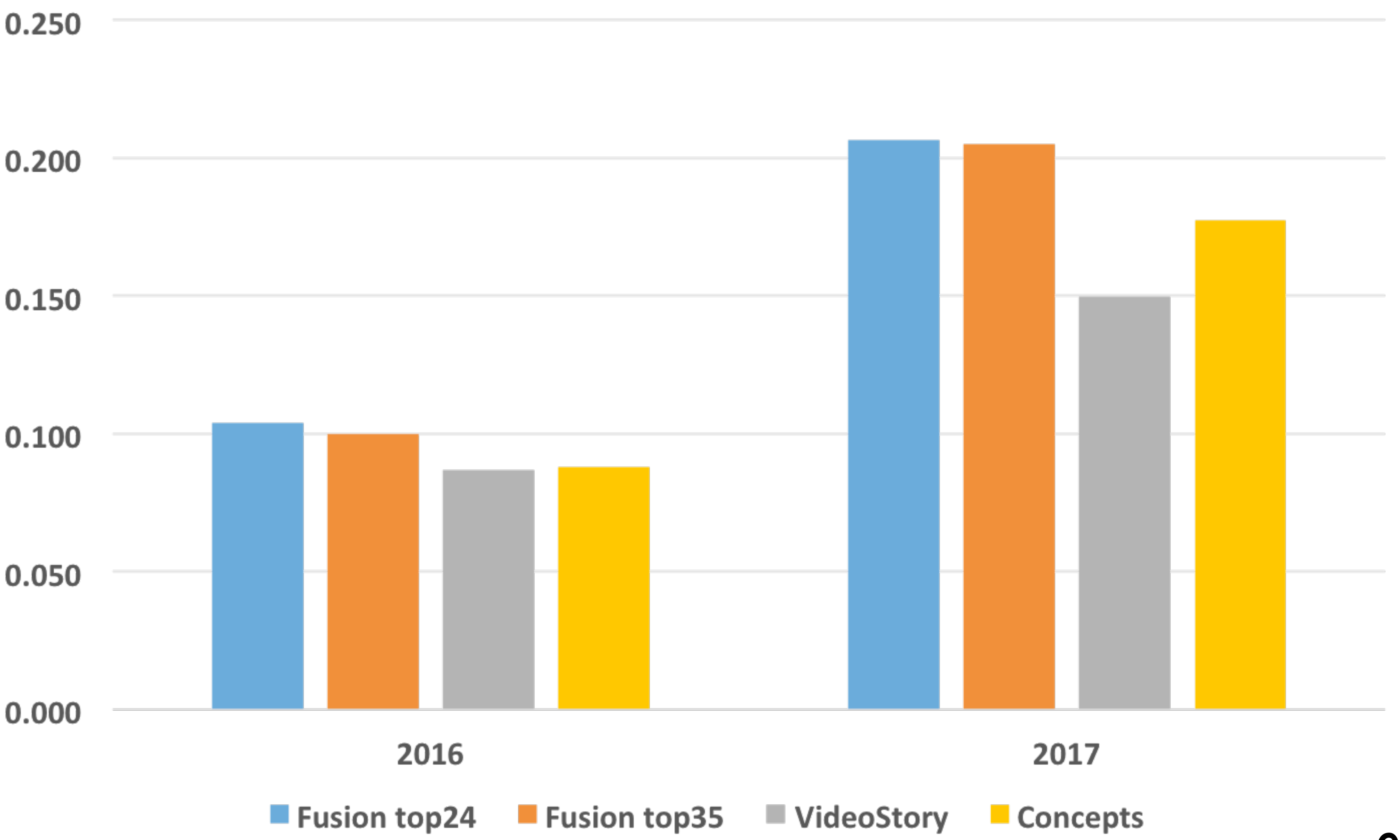
Final Fusion

- Concept fusion is slightly better than VideoStory
- Often complementary, also big difference for many topics
- Top 2/4 for concepts is slightly better than top 3/5



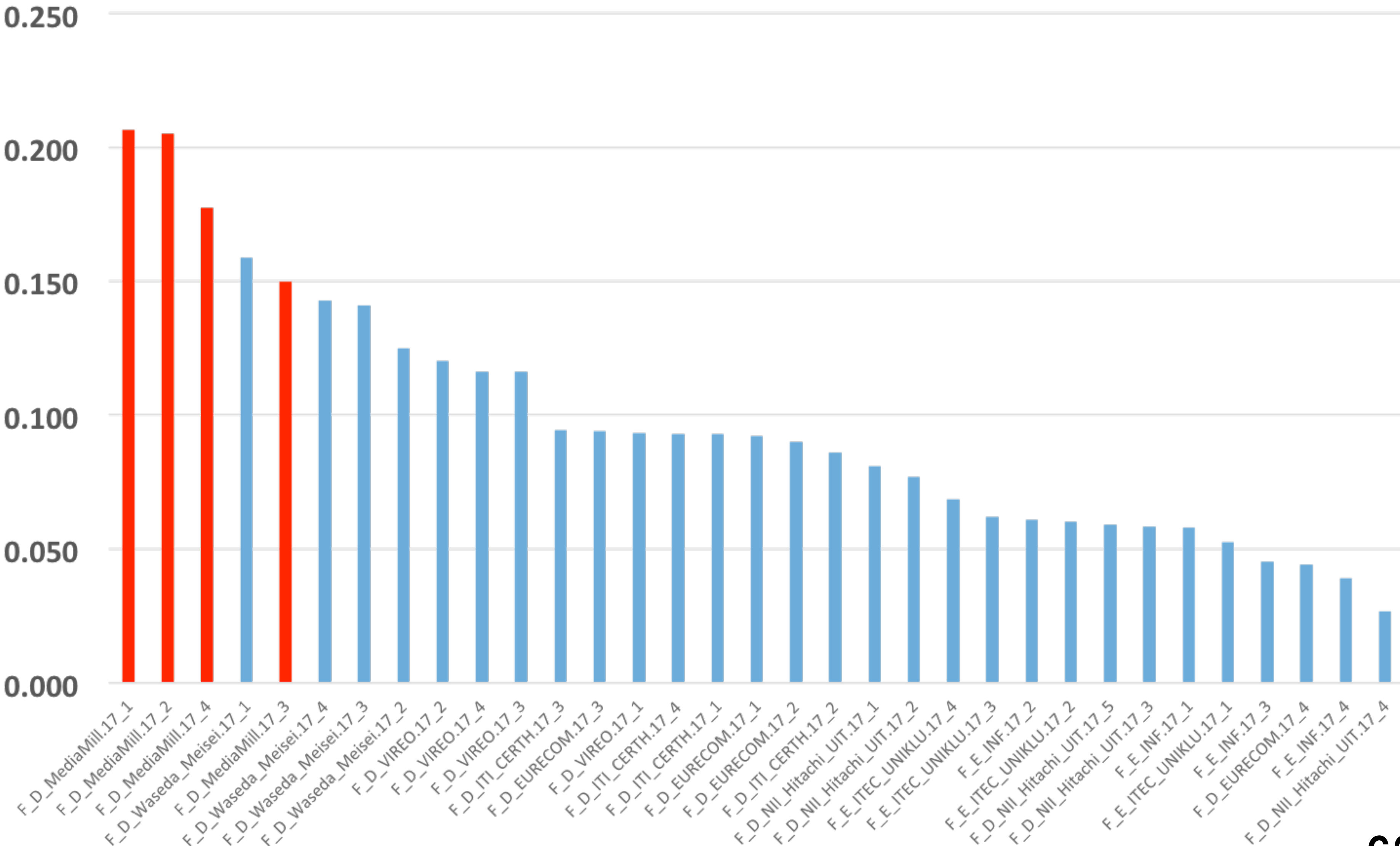
MediaMill system

AVS Submission



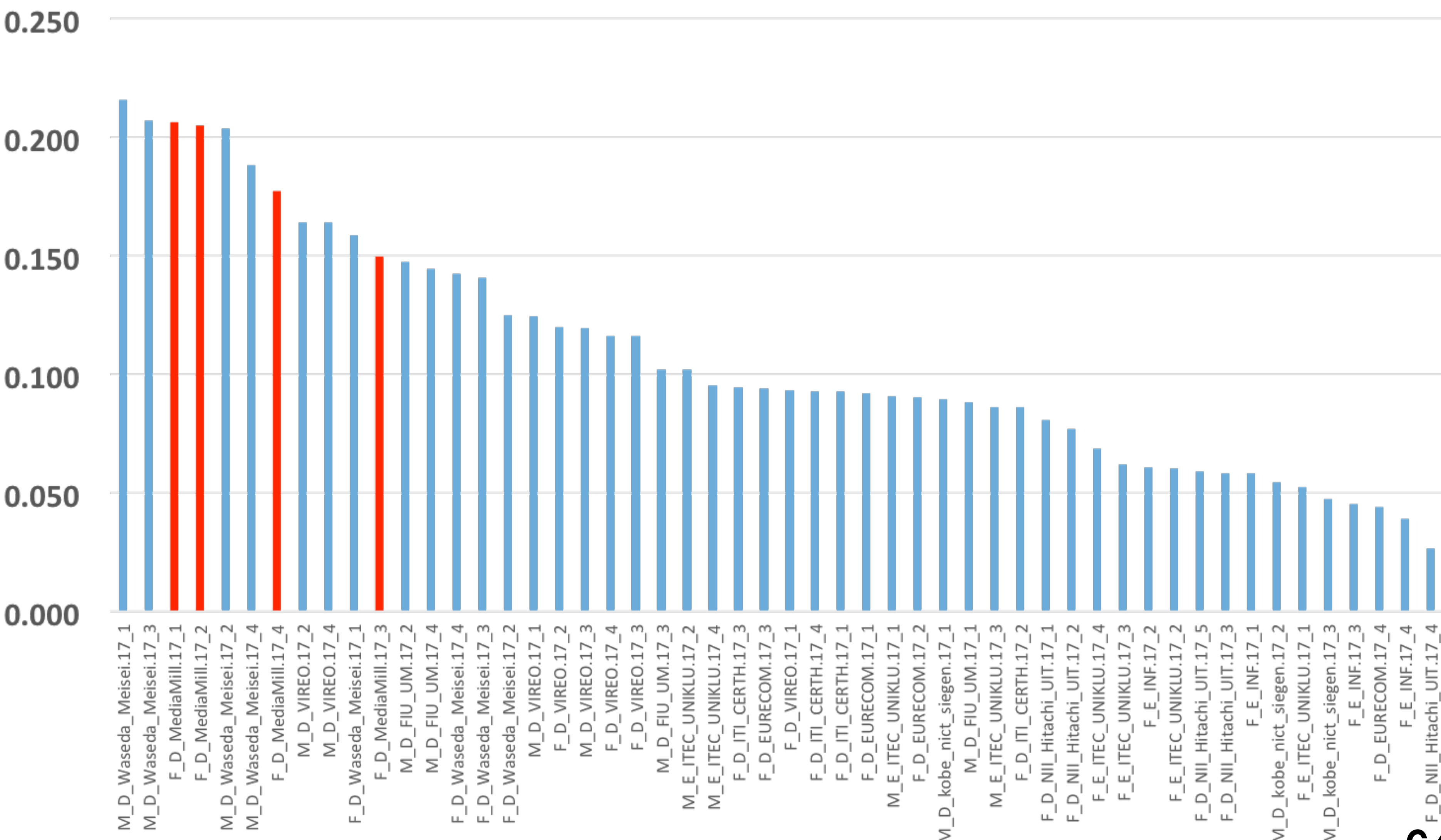
MediaMill system

All Fully Automatic AVS Submissions



MediaMill system

All Automatic and Interactive AVS Submissions



Conclusions

- Query parsing is important
- VideoStory and Concepts are good but will not “solve” AVS

Part III:
Summary and future works

Summary

2017 main approaches

- Concept bank with automatic or manual mapping with query terms
- Combination of concept scores from Boolean operators
- Work on Query Understanding
- Rectified Linear Score Normalization
- Use of Video-To-Text techniques on shots
- Query expansion / term matching techniques
- Use of unified text-image vector space

TREC Video Retrieval Evaluation Notebook Papers and Slides

https://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html

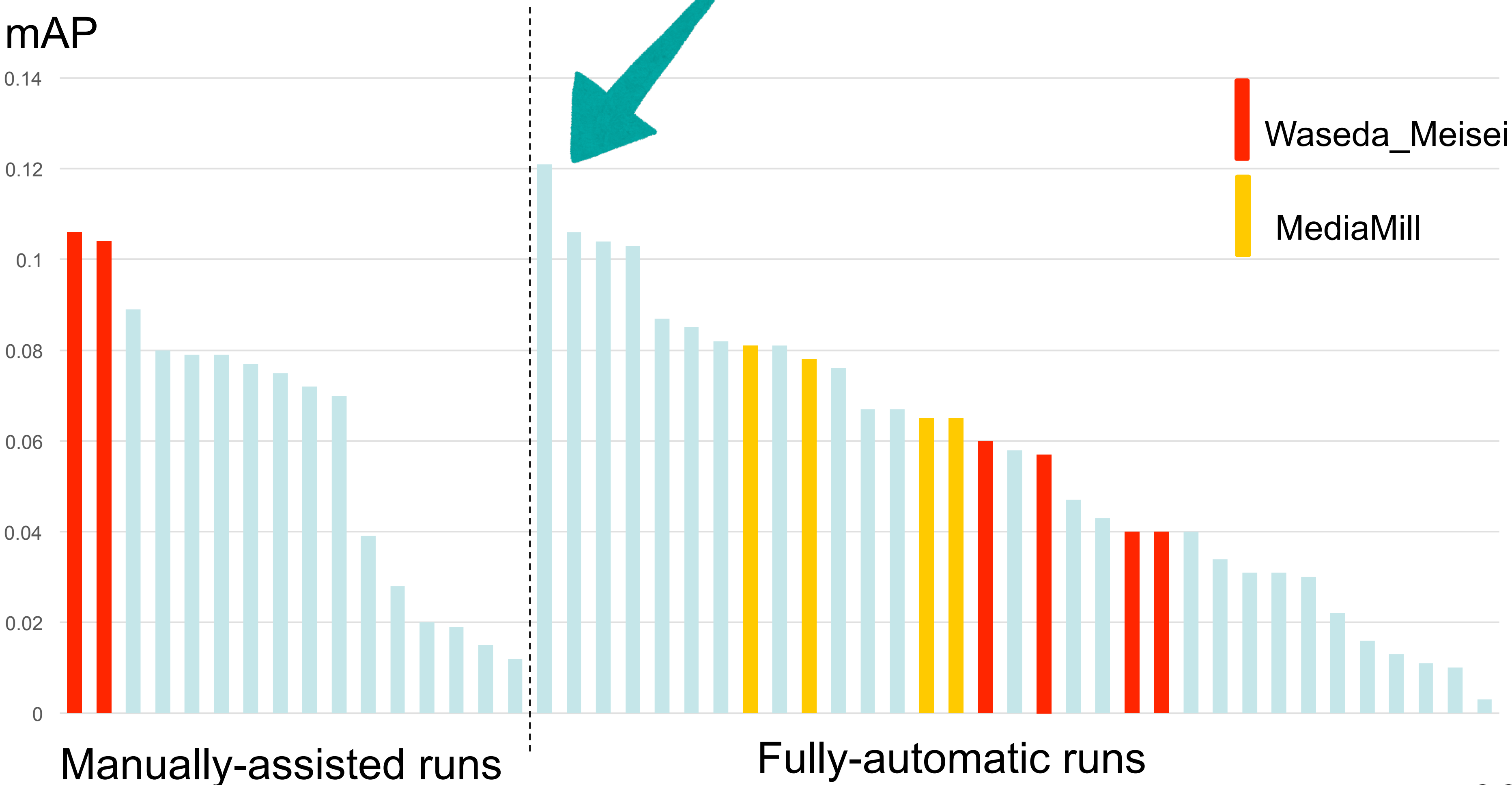
Summary

2017 observations

- Ad-hoc search is more difficult than simple concept-based tagging.
- Max and Median scores are better than TRECVID 2016 for both manually-assisted and fully-automatic runs
- Manually-assisted runs performed slightly better than automatic.
- Most systems are not real-time (slower systems were not necessarily effective)

TRECVID 2018 results NEW

Some of the fully-automatic systems performed better than the concept-bank based manually assisted system!



Future works

- Concept bank based methods are good but will not be able to solve “AVS” task.
- Comprehend query phrases linguistically and utilize more human knowledge.
- Directly search for videos without decomposing the query.

We will discuss more about this task and new approaches at TRECVID workshop on 13 - 15 Nov.

We are waiting for new participants next year!