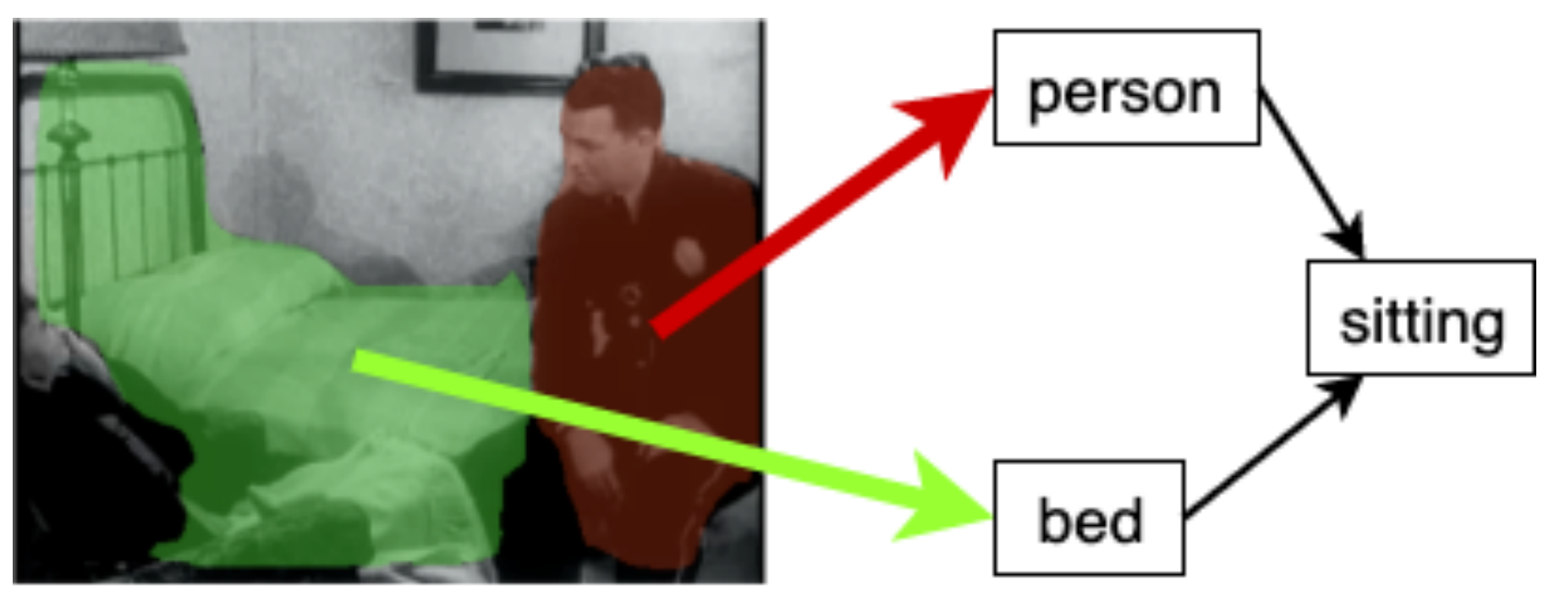
Kindai University and Kobe University at TRECVID 2019 AVS Task

Kimiaki Shirahama[1], Daichi Sakurai[1], Takashi Matsubara[2], Kuniaki Uehara[2]
1. Department of Informatics, Kindai University, 2. Graduate School of System Informatics, Kobe University

Motivation

- **Concept-based approach**
  1. Numerous concepts.
  2. Exponentially increasing number of concept relations

  → Impossible to prepare models that detect all the concepts and relations.

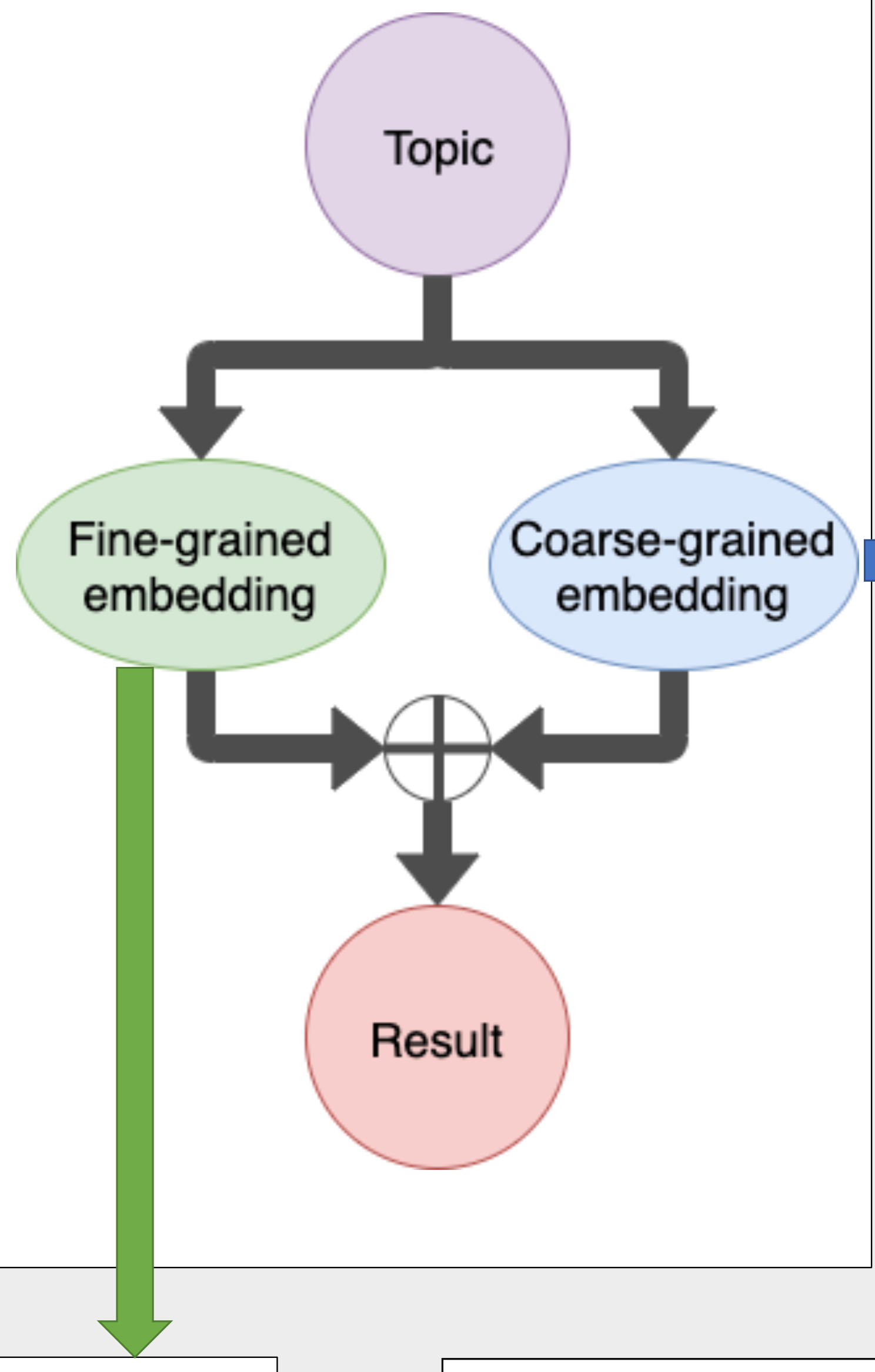


A person sitting on the bed

- **Embedding approach**

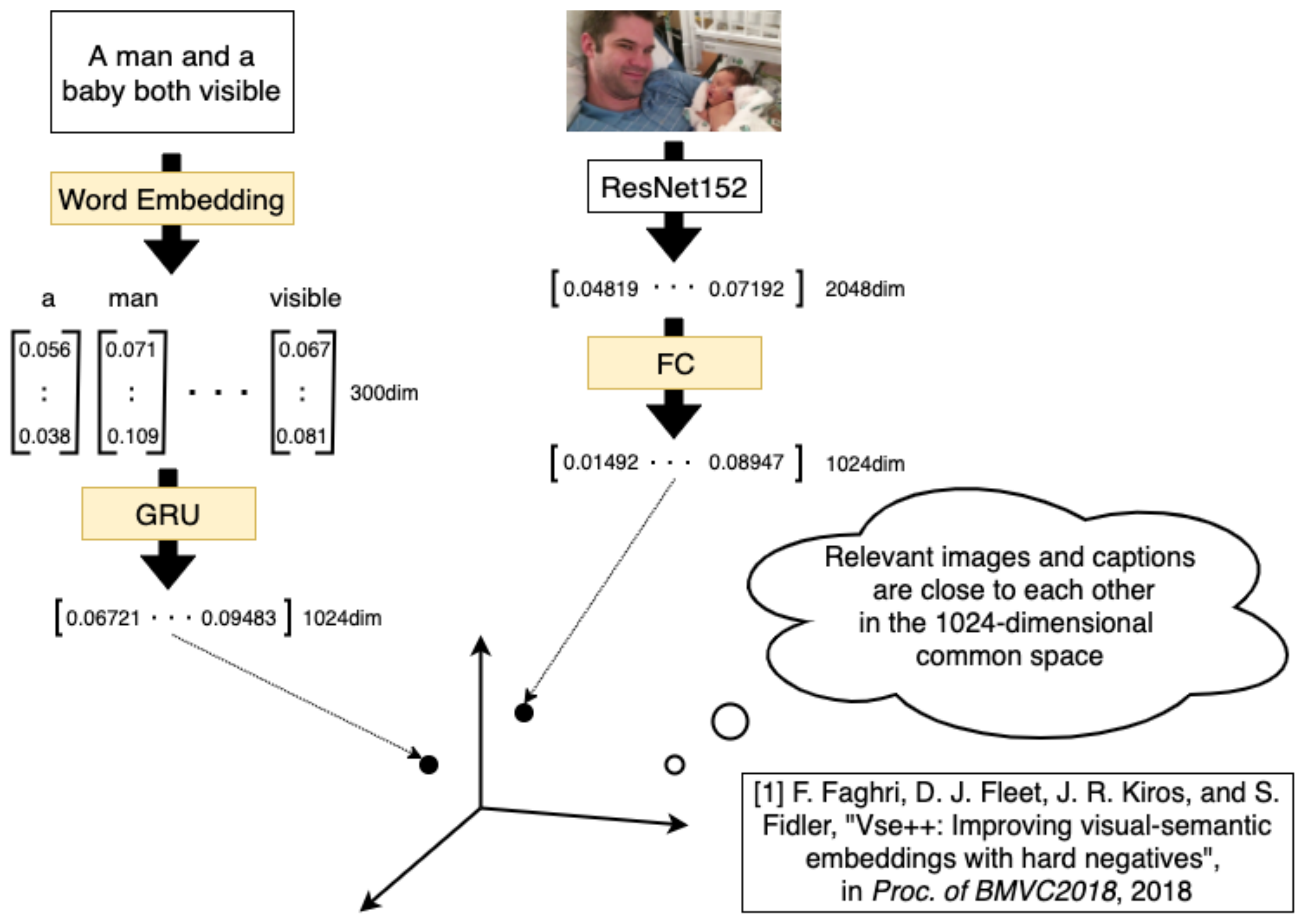
  Map visual features of a shot and textual features of a topic into a common space.
  → Their similarity can be directly computed.

Overview



Coarse-grained embedding (VSE++)

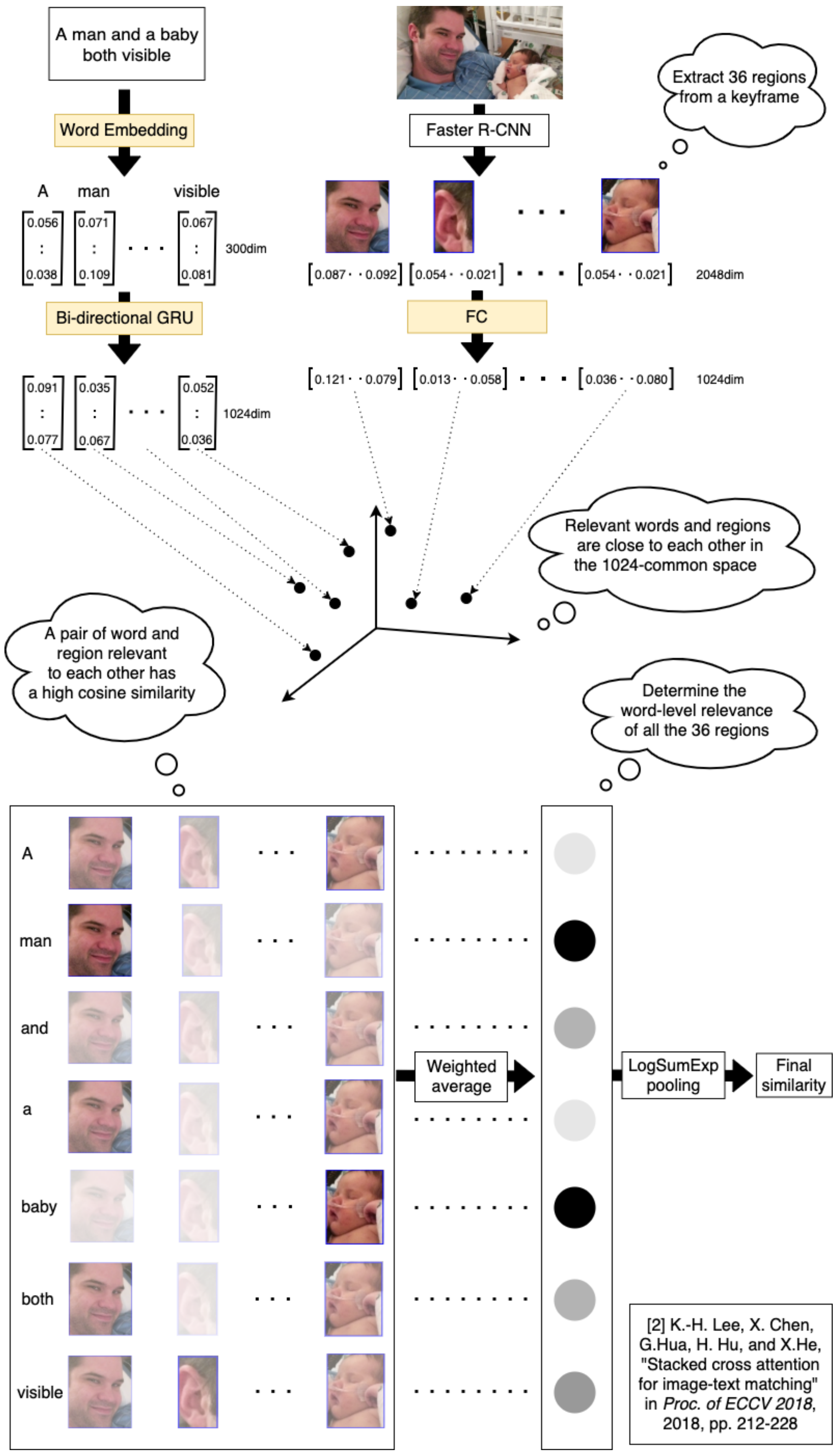
- Map frames in a shot and the text description of a topic into a common space.
- It is useful to evaluate the overall relevance of a shot to the topic.


A man and a baby both visible
Word Embedding
a
man
visible
0.056
0.038
0.071
0.109
0.067
0.081
300dim
GRU
[0.06721 · · · 0.09483] 1024dim
ResNet152
[0.04819 · · · 0.07192] 2048dim
FC
[0.01492 · · · 0.08947] 1024dim
Relevant images and captions are close to each other in the 1024-dimensional common space

[1] F. Faghri, D. J. Fleet, J. R. Kiros, and S. Fidler, "Vse++: Improving visual-semantic embeddings with hard negatives", in *Proc. of BMVC2018*, 2018

Fine-grained embedding (SCAN)

- Build a common space to characterize correspondences between regions in the keyframe of a shot and words in the text description of a topic.
- It is useful to examine whether a shot satisfies detailed requirements of a topic, such as object numbers, object types and object characteristics.


A man and a baby both visible
Word Embedding
A
man
visible
0.056
0.038
0.071
0.109
0.067
0.081
300dim
Bi-directional GRU
0.091
0.077
0.035
0.067
0.052
0.036
1024dim
Faster R-CNN
Extract 36 regions from a keyframe
[0.087· · 0.092] [0.054· · 0.021] [0.054· · 0.021] 2048dim
FC
[0.121· · 0.079] [0.013· · 0.058] [0.036· · 0.080] 1024dim
Relevant words and regions are close to each other in the 1024-common space
A pair of word and region relevant to each other has a high cosine similarity
Determine the word-level relevance of all the 36 regions
A
man
and
a
baby
both
visible
Weighted average
LogSumExp pooling
Final similarity

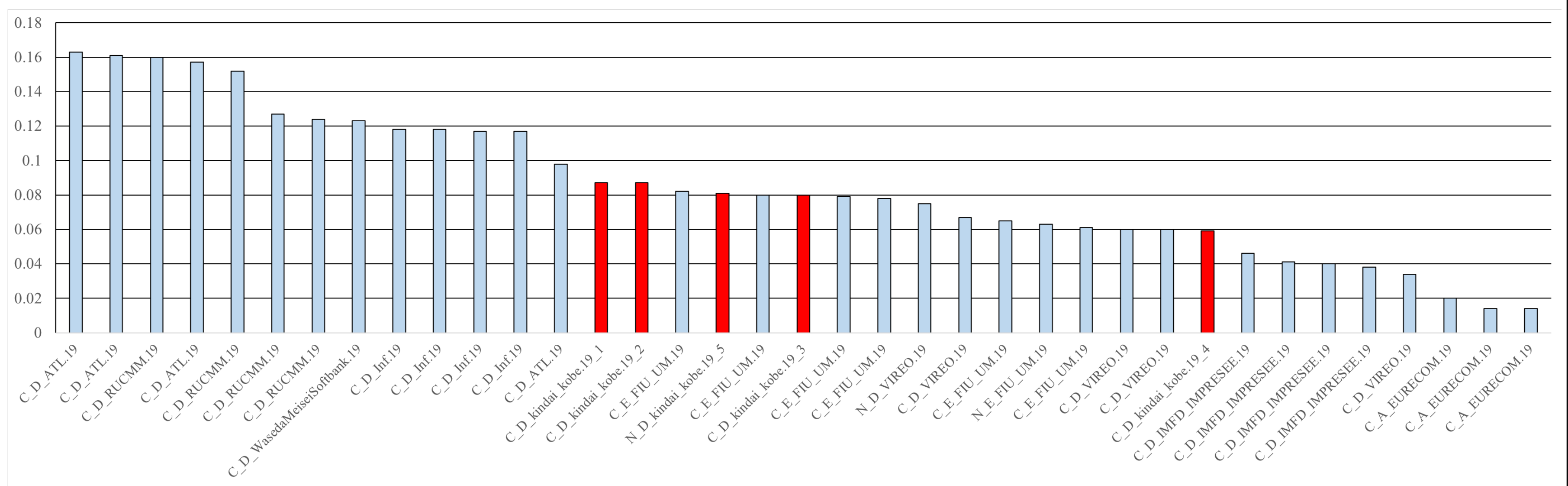
[2] K.-H. Lee, X. Chen, G.Hua, H. Hu, and X.He, "Stacked cross attention for image-text matching" in *Proc. of ECCV 2018*, 2018, pp. 212-228

Future work

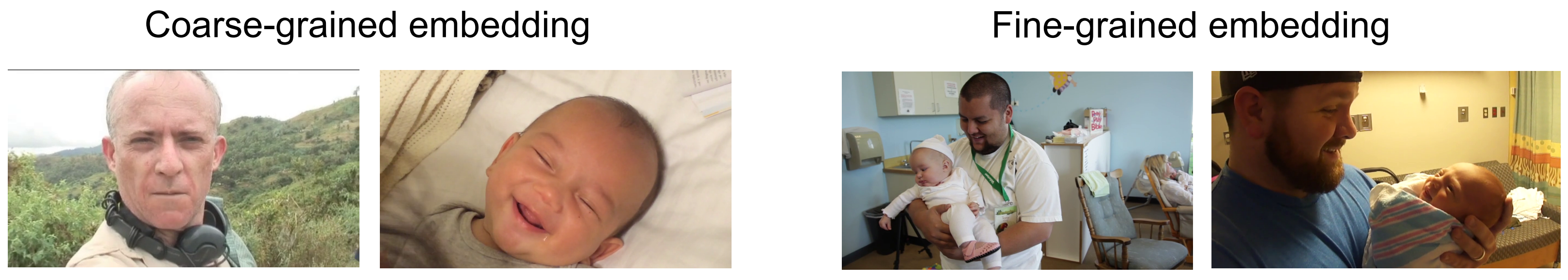
- Develop a fast matching method that efficiently filters out many irrelevant shots by considering the structure of a topic's text description and the relation among regions.
- Exploit Web images annotated with words which are not included in captions of dataset's vocabulary.

Results

i. C_D_kindai_kobe.19_1 is an ensemble of VSE++M, VSE++F and SCAN
   VSE++M trained on MS-COCO, VSE++F trained on Flickr 30k, SCAN trained on MS-COCO
ii. C_D_kindai_kobe.19_2 is an ensemble of VSE++M, VSE++F and SCAN where the feature of each region is L2-normalised
iii. C_D_kindai_kobe.19_3 is comprised only of SCAN
iv. C_D_kindai_kobe.19_4 is an ensemble of VSE++M and VSE++F
v. N_D_kindai_kobe.19_5 is comprised only of SCAN with L2normalization



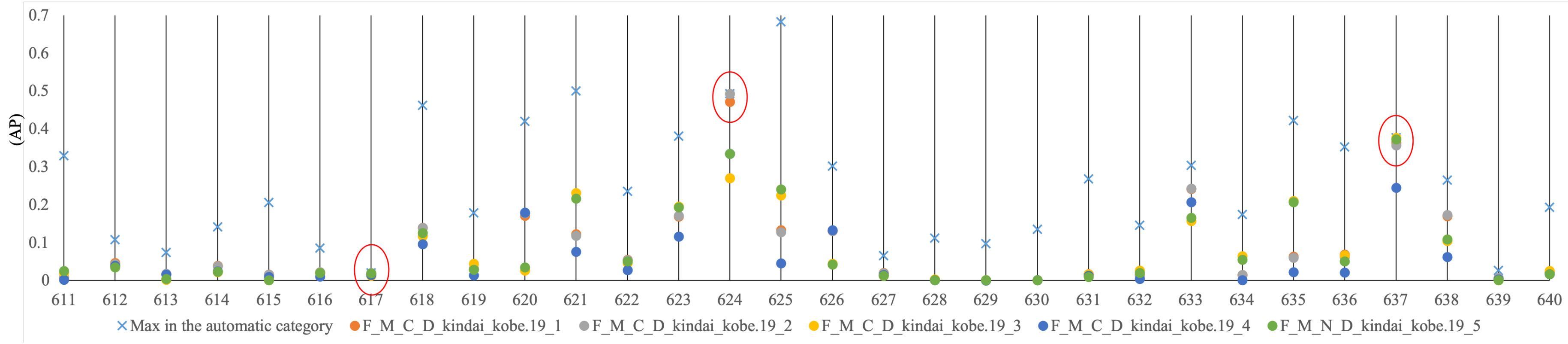
1. C_D_kindai_kobe.19_1 and C_D_kindai_kobe.19_2 are ranked at the fifth position in terms of teams participating in the fully-automatic category.
2. The ensemble of VSE++ and SCAN leads to a performances improvement.
3. Fine-grained embedding based on SCAN is much more effective than coarse grained embedding based on the ensemble of VSE++M and VSE++F.



Topic 636: A man and a baby both visible

Many shots only including man's or baby's appearance are retrieved.

Shots including both man's and baby's appearances are retrieved.



Our runs achieve the best APs for three topics in the fully-automatic category.

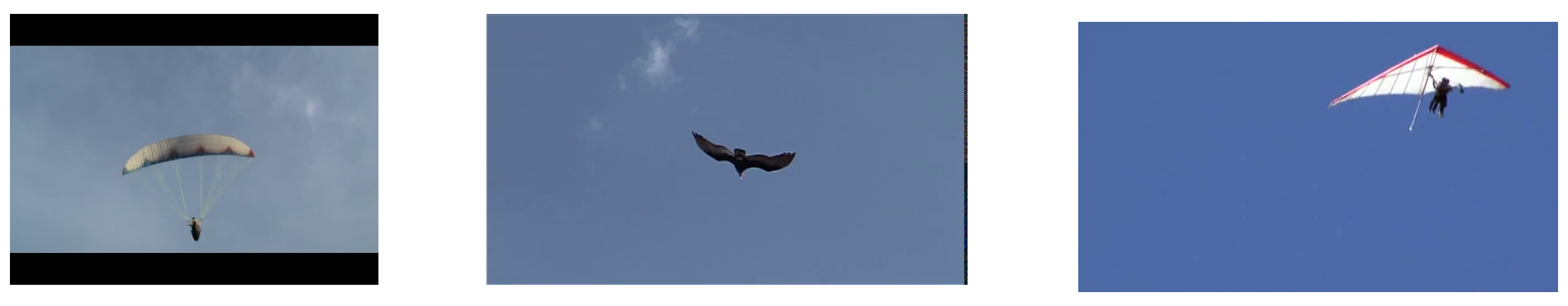
Analysis on failure cases

- For topics involving phrases

Topic 629: black man singing

This caption should be split into "black man" and "singing", but it is split into "black", "man" and "singing".
→ Many retrieved shots show black background or clothes and not black man who singing

- For topics involving words which are not included in dataset's vocabulary



Topic 611: a drone flying

"Drone" exists in neither MS-COCO's nor Flicker30k's vocabulary.
→ Many retrieved shots show a flying bird and a person who is parachuting or hang-gliding