Word2VisualVec++ for Ad-hoc Video Search

Xirong Li[1], Chaoxi Xu[1], Jianfeng Dong[2], Jing Cao[1],
Xun Wang[2], Yang Gang[1]

[1]**Renmin University of China**

[2]Zhejiang Gongshang University

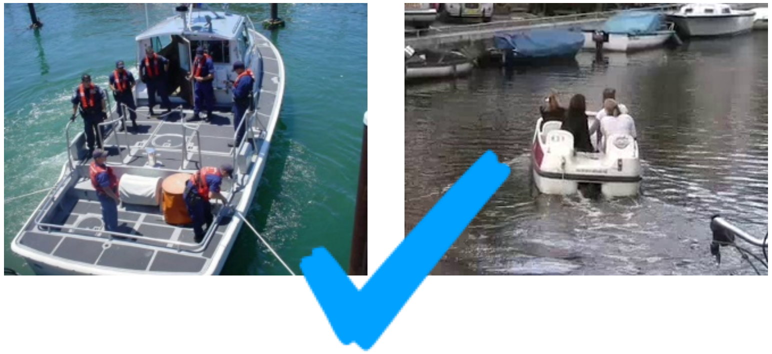
TRECVID 2018 Workshop
2018-11-13

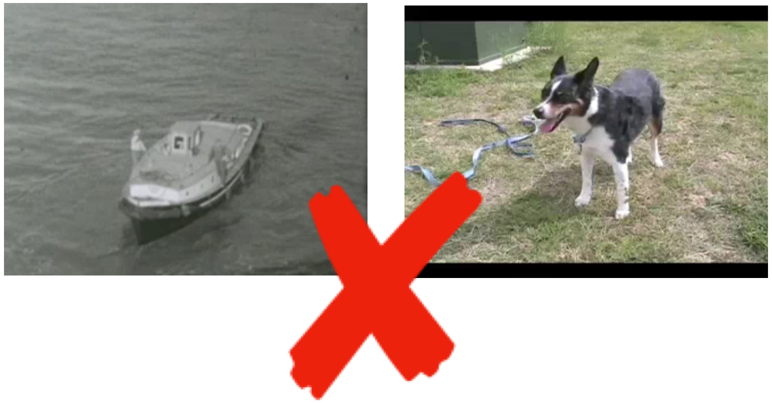
Task: Ad-hoc Video Search

A natural-language query, no visual example provided

- This is ***zero-shot*** video retrieval

Find shots of one or more people on a moving boat in the water





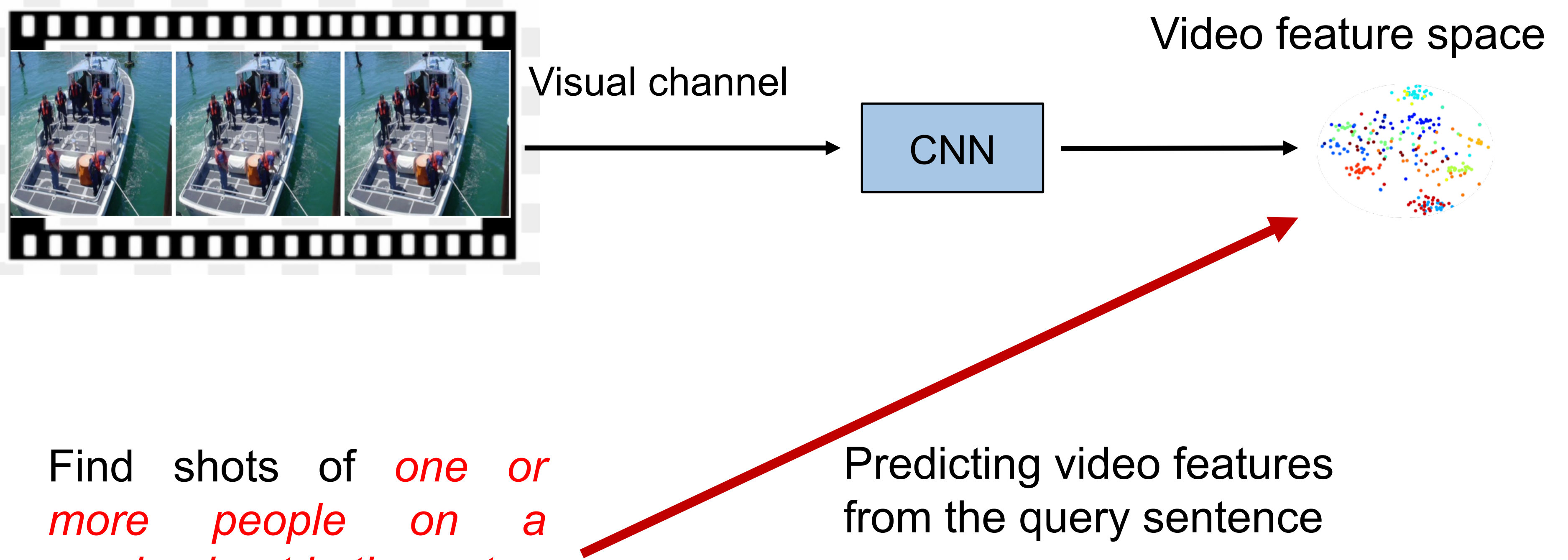
Compute cross-modal video-text similarity given **two challenges**

1. **Ad-hoc** means **open vocabulary**
2. **Natural text** requires **sequential modeling**

Our Idea

Compute video-text similarity in a **video feature space**

- As we did in TV16 / TV17 for the VTT task



Visual channel

CNN

Video feature space

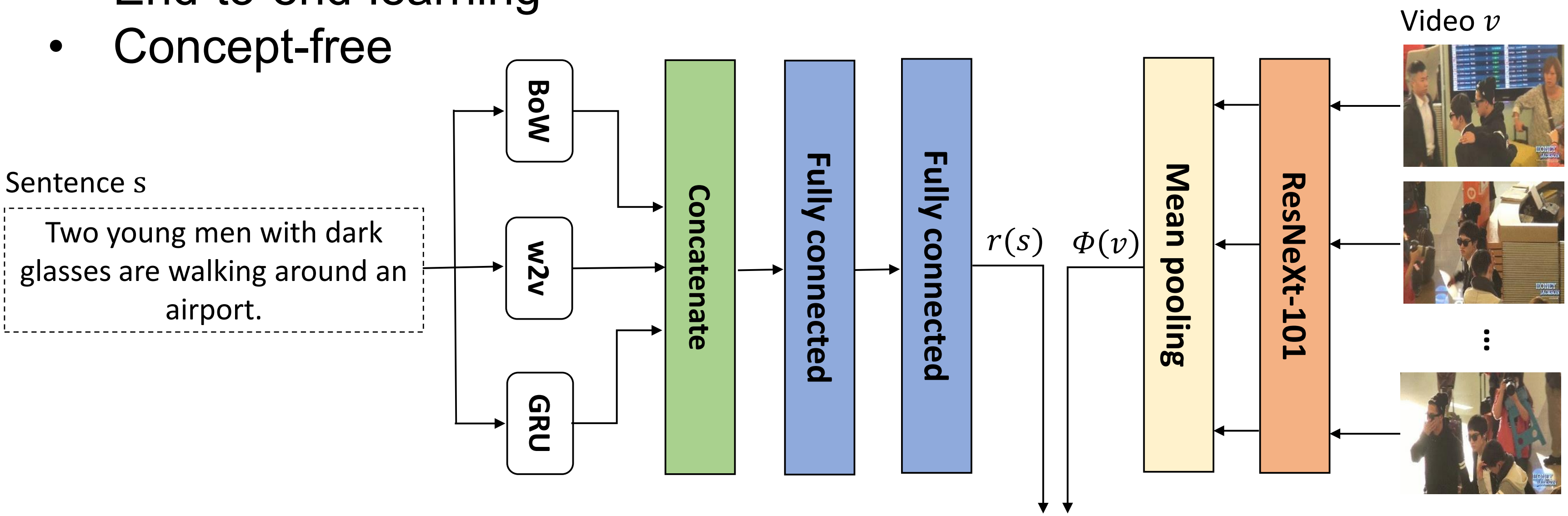
Find shots of *one or more people on a moving boat in the water*

Predicting video features from the query sentence

Our Solution

Build on the top of the Word2VisualVec model

- End-to-end-learning
- Concept-free



$(r(s) - \Phi(v))^2$

Minimize mean squared error

Multi-scale sentence encoding

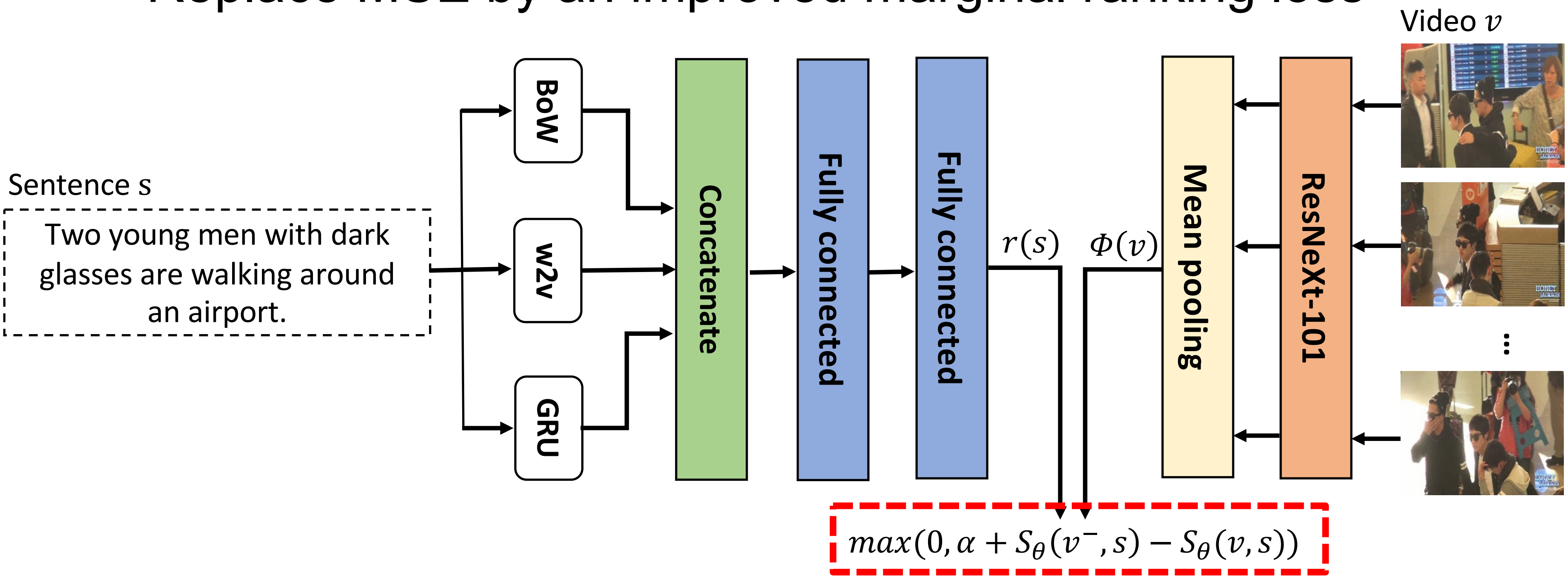
- word2vec for open vocabulary
- GRU for sequential modeling

Dong et al., Predicting Visual Features from Text for Image and Video Caption Retrieval, T-MM 2018

Our Solution

Word2VisualVec → **Word2VisualVec++**

- Replace MSE by an improved marginal ranking loss


Video v
Sentence s
Two young men with dark glasses are walking around an airport.
BoW
w2v
GRU
Concatenate
Fully connected
Fully connected
$r(s)$
$\Phi(v)$
Mean pooling
ResNeXt-101
$max(0, \alpha + S_\theta(v^-, s) - S_\theta(v, s))$

v^- denotes the hardest negative video sample of the sentence s

Faghri et al., VSE++: Improving visual-semantic embeddings with hard negatives, BMVC 2018

Our Solution

Dataset	Usage	No. videos	No. frames
msrvtt10k	training	10,000	305,462
tgif	training	100,855	1,045,268
TV16 VTT training set	validation	200	5,941

Frame-level features	Dim.
ResNext-101	2,048
ResNet-152	2,048

https://github.com/li-xirong/avs

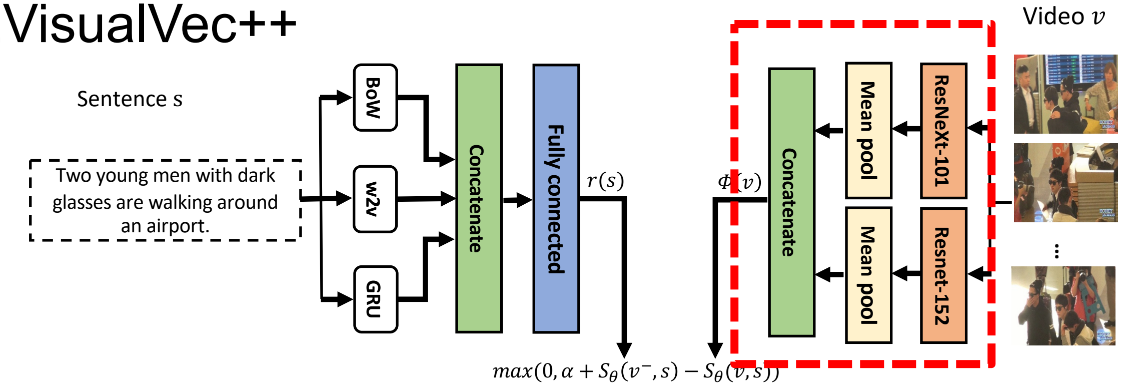
Snoek et al., University of Amsterdam and Renmin University at TRECVID 2017, TRECVID 2017
Dong et al., DI-61-86 at TRECVID 2017: Video-to-text description. TRECVID 2017

Our Solution

Three variants of Word2VisualVec++

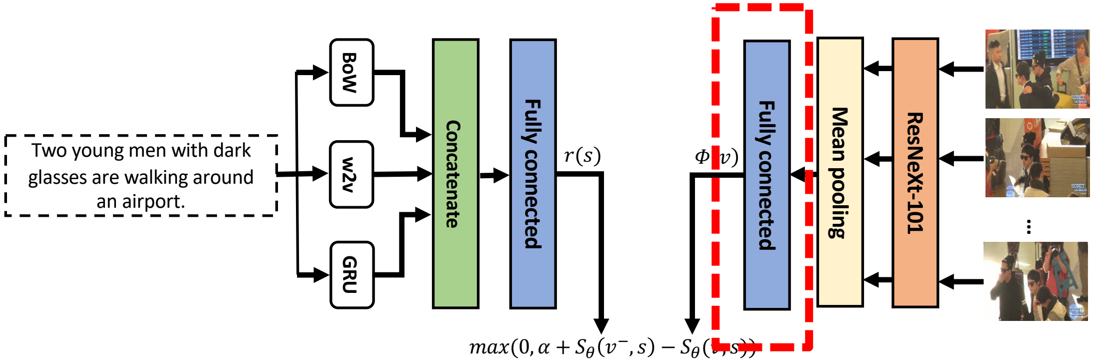
(1) Model for *Run 4*

Feature concatenation



(2) Model for *Run 3*

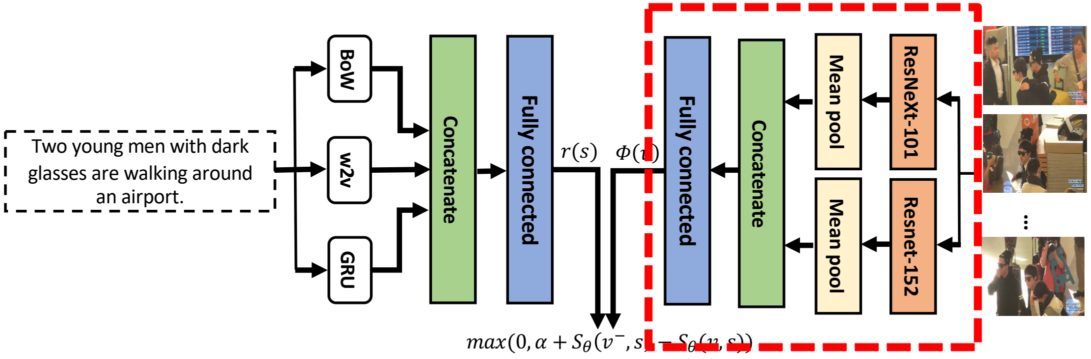
Feature re-learning


Two young men with dark glasses are walking around an airport.
BoW
w2v
GRU
Concatenate
Fully connected
$r(s)$
$\Phi(v)$
Fully connected
Mean pooling
ResNeXt-101
$max(0, \alpha + S_\theta(v^-, s) - S_\theta(v, s))$

(3) Model for *Run 2*

Feature concatenation

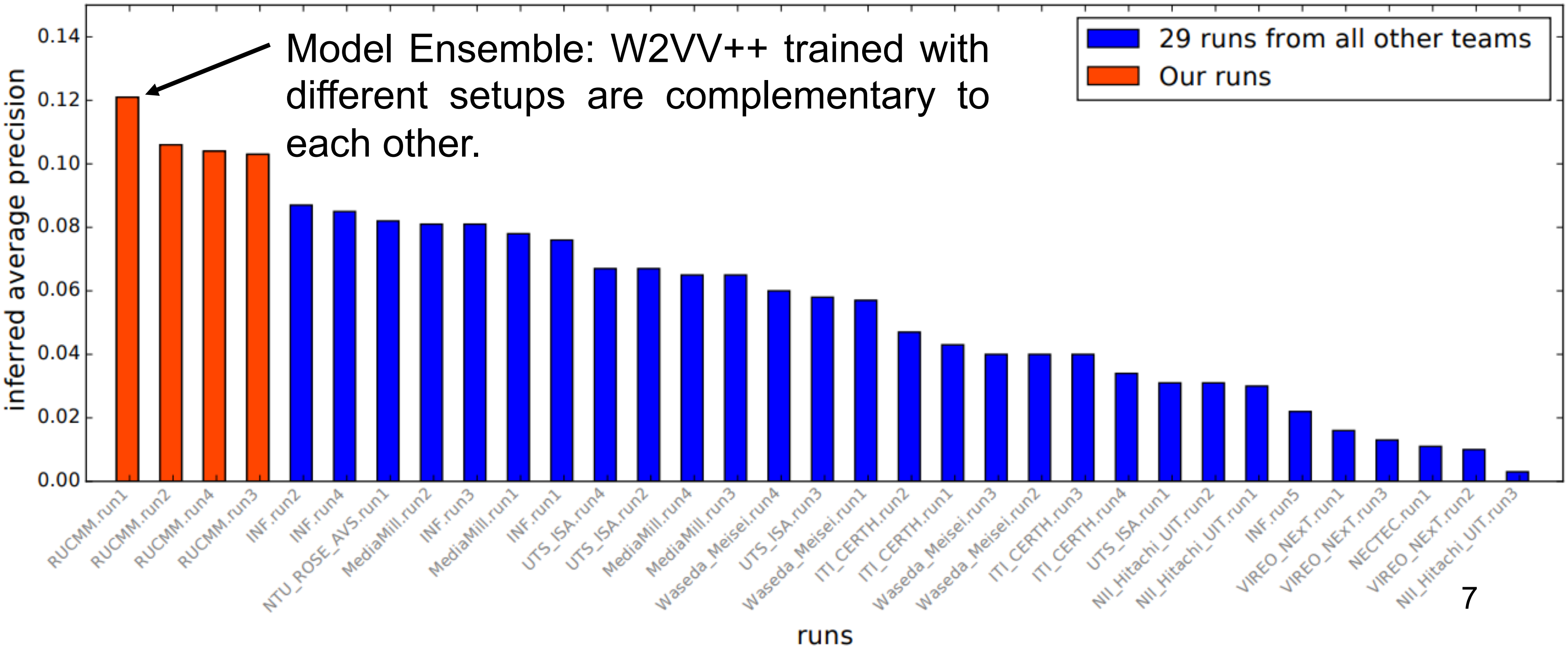
Feature re-learning



Overall Evaluation Results

Our submissions top the performance.

- Run 1 equally combines multiple W2VV++ trained with different setups.
- Run 1 > Run 2 > Run 4 > Run 3



Model Ensemble: W2VV++ trained with different setups are complementary to each other.

Results of individual topics

Topic	Run4	Run3	Run2	Run1
561	0.049	0.039	0.114	0.080
562	0.066	0.076	0.06	0.087
563	0.456	0.422	0.511	0.492
564	0.158	0.178	0.224	0.205
565	0.247	0.389	0.319	0.319
566	0.046	0.036	0.041	0.067
567	0.011	0.005	0.012	0.009
568	0.068	0.087	0.069	0.075
569	0.017	0.01	0.018	0.022
570	0.000	0.011	0.002	0.010
571	0.090	0.103	0.118	0.096
572	0.046	0.078	0.085	0.137
573	0.089	0.179	0.172	0.235
574	0.057	0.02	0.007	0.051

Topic	Run4	Run3	Run2	Run1
575	0.032	0.059	0.060	0.156
576	0.004	0.005	0.027	0.008
577	0.343	0.325	0.056	0.381
578	0.323	0.033	0.127	0.011
579	0.063	0.030	0.026	0.020
580	0.011	0.004	0.027	0.005
581	0.226	0.229	0.213	0.249
582	0.007	0.016	0.008	0.020
583	0.152	0.069	0.192	0.177
584	0.292	0.296	0.315	0.301
585	0.177	0.240	0.271	0.275
586	0.043	0.054	0.037	0.057
587	0.006	0.010	0.014	0.018
588	0.031	0.026	0.037	0.044
589	0.015	0.052	0.027	0.023
590	0.005	0.002	0.003	0.002

Seven topics with infAP < 0.02

Case study

567 Find shots of people performing or dancing outdoors at nighttime (infAP: 0.009)

Top-10 results

shot37195_365_4951

shot37195_305_4752

shot37195_362_4941

shot37195_304_4748

shot37195_318_4795

shot37195_313_4779

shot37195_328_4829

shot37195_329_4832

shot37195_309_4766

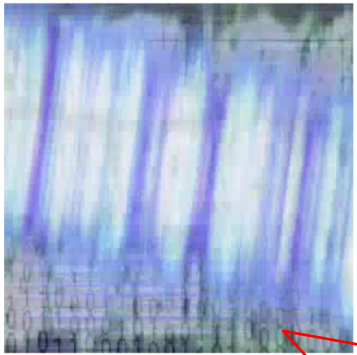
shot37195_346_4888

The top ranked results seem correct ☺

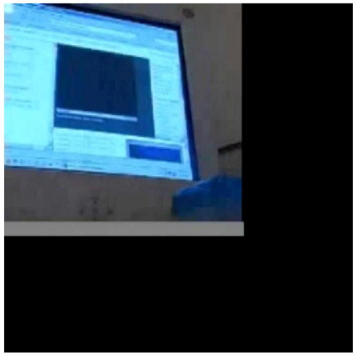
Case study

570 Find shots of a projection screen (infAP: 0.010)

Top-5 results

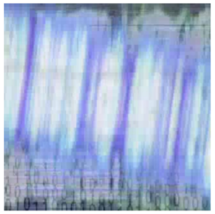

shot35694_19_1521

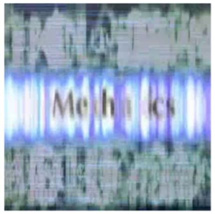
shot36510_242_3694


shot38408_44_4410

shot35694_84_7576

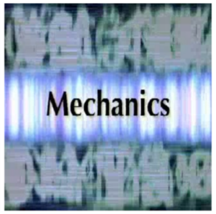
shot36510_383_5715


shot35694_19_1521

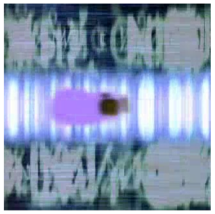


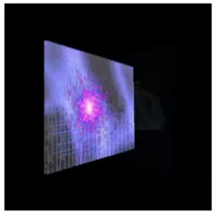
shot35694_19_1535

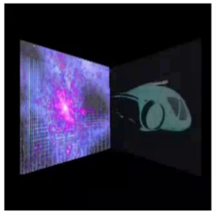
shot35694_19_1549

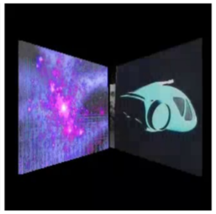


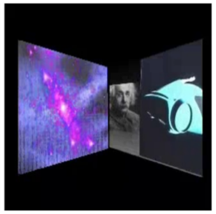
shot35694_19_1563


shot35694_19_1577


shot35694_19_1591


shot35694_19_1605


shot35694_19_1619


shot35694_19_1634

shot35694_19_1648

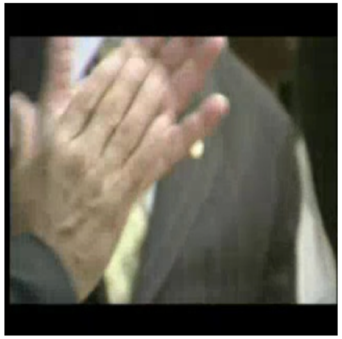
Looks like a projected screen ☺

Case study

576 Find shots of a person holding his hand to his face (infAP: 0.008)

Top-10 results

shot35673_21_1472



shot38899_56_3928

shot36772_52_3590

shot38814_67_5006

shot36772_56_3637

shot35673_15_1424

shot36772_54_3615

shot38875_94_8696

shot38334_75_12755

shot37193_462_12373

“Face” seems to be ignored ☹

Retrospective experiments

We used our TV18 system, as is, to answer TV16 / TV17 AVS topics.

Run	TV16	TV17	TV18
Previous best run	0.054 [A]	0.206 [B]	-
Our TV18 Runs:			
Run 4	0.149	0.176	0.104
Run 3	0.140	0.171	0.103
Run 2	**0.151**	0.213	0.106
Run 1	0.149	**0.220**	**0.121**

Topic difficulty: TV18 > TV16 > TV17

[A] Le et al., NII-HITACHI-UIT at TRECVID 2016, TRECVID 2016
[B] Snoek et al., University of Amsterdam and Renmin university at TRECVID 2017, TRECVID 2017

Conclusions & Observations

Word2VisualVec++ is quite effective for the AVS task

- Top performer for TV16 / 17 / 18

Model ensemble is a good trick

- Improve infAP from 0.106 (single model) to 0.121

Concept-free can be a double-edged sword

- Results might be less interpretable than concept-based methods
- An interesting direction to pursue.