VIREO-EURECOM @ TRECVID 2019: Ad-hoc Video Search (AVS)

Phuong Anh Nguyen[†], Jiaxin Wu[†], Chong-Wah Ngo[†], Francis Danny[⋆], Benoit Huet[⋆]

[†] *Video Retrieval Group (VIREO), City University of Hong Kong*
http://vireo.cs.cityu.edu.hk

[⋆] *Data Science Department, EURECOM, France*
http://www.eurecom.fr

Abstract

In this paper, we describe the systems developed for Ad-hoc Video Search (AVS) task at TRECVID 2019[1] and the achieved results.

Ad-Hoc Video Search (AVS): We merge three video search systems for AVS, including: two concept-based video search systems which analyse the query using linguistic approaches then select and fuse the concepts, and a video retrieval model which learns the joint embedding space of the textual queries and the videos for matching. With this setting, we plan to analyze the advantages and shortcomings of these video search approaches. We submit totally seven runs consisting four automatic runs, two manual runs, and one novelty run. We brief our runs as follows:

- *F_M_C_D_VIREO.19_1*: This automatic run has *mean xinfAP=0.034* using a concept-based video search system including $\sim$16.6k concepts covering objects, persons, activities, and places. We parse the queries with Stanford NLP parsing tool [2], keep the keywords, and categorize the keywords into three groups: object/person, action, and place. Correspondingly, the concepts from different groups in the concept bank are selected and fused.
- *F_M_C_D_VIREO.19_2*: This automatic run has *mean xinfAP=0.067* using the second concept-based video search system including $\sim$16.4k concepts. The concept bank is slightly different comparing to the concept bank used in *F_M_C_D_VIREO.19_1*. From the query, we embed the words, terms, and the whole query by the Universal Sentence Embedding [3]. Similarly, we use the same method to embed all the concept names in the concept bank. Finally, the concepts are selected by an incremental concept selection method [4] based on the cosine similarity of the embedded query and embedded concept name.
- *F_M_C_D_VIREO.19_3*: This run is a fusion of the results from three different automatic runs: *F_M_C_D_VIREO.19_1*, *F_M_C_D_VIREO.19_2*, and *F_M_C_A_EURECOM.19_1*. In the run *F_M_C_A_EURECOM.19_1*, three embedding spaces are learnt separately for object counting, activity detection, and semantic concept annotation. The query textual feature and the video visual feature are mapped into these three embedding spaces and fused for matching. The run ends up with *mean xinfAP=0.060*.
- *F_M_C_D_VIREO.19_4*: This run is a fusion of the results from three automatic runs mentioned in the run *F_M_C_D_VIREO.19_3* together with the result of a metadata-based retrieval system.

To enable metadata search, we index all the video metadata by Lucene and the retrieval is done in video level. The performance stays at *mean xinfAP=0.060.*

- *M_M_C_D_VIREO.19_1*: This manual run uses the same system with the same settings presented in the run *F_M_C_D_VIREO.19_1.* The difference is that the user parses and categorizes the query manually at the beginning of the process. This human intervention improves the *mean xinfAP* from *0.034* to *0.066.*
- *M_M_C_D_VIREO.19_2*: This manual run uses the same system with the same settings presented in the run *F_M_C_D_VIREO.19_2.* After getting the list of selected concepts for each query, the user screens the concept list and remove unrelated or unspecific concepts to refine the result. This step helps improving the *mean xinfAP* from *0.067* to *0.118.*
- *F_M_N_D_VIREO.19_5*: This is the novelty run with *mean xinfAP=0.075*, and this is the best automatic run from VIREO team. The system used to process the query is in the same settings with the system presented in the run *F_M_C_D_VIREO.19_2* except that we only use the embedding of the whole query sentence for concept selection.

1 Ad-Hoc Video Search (AVS)

1.1 Detail descriptions

1.1.1 F_M_C_D_VIREO.19_1

The main idea of this run is to compare concept in three kinds of aspects including object/person, action and place. Firstly, we extract these three kinds of concepts from video shots and divide the query sentence into these three aspects. Then we compute the concept similarity in each aspect to get three individual similarities. Finally, we compute the whole similarity of the sentence and video shots by combing these three individual similarities together. The final ranking list depends on the combined similarity.

Correspondingly, we have three types of concept bank. Table 1 shows the details of our concept bank and the models we used in each dataset. Using this concept bank, we calculate all concept scores for all video shots in three aspects by adopting the pre-trained classification models. If the concept score is higher than the threshold θ, then it is regarded that the video shot has this concept. In this run, θ is set to 0.1.

For the query sentence, we also divided it into three aspects. We first use Stanford NLP parsing tool [2] to parse the sentence and get the dependency of the words and their parts of speech. Then we extract keywords (key phases) from the sentence and classify them into three categories (object/person, action and place) based on the parts of speech and dependency. For example, given a query "a crowd of people attending a football game in a stadium", "a crowd of people" is classified as the object/person category, "attending a football game" is classified as the action category, and "in a stadium" is classified as the place category.

After extracting concepts from video shots and keywords (key phrases) from the query sentence, we compute the text similarity between them by using a word2vec model [5]. Finally, given a sentence query q, the similarity score s_i for a video shot v_i is computed as follows:

$$s_i = w_{object} \times sim_{object}(q, v_i) + w_{action} \times sim_{action}(q, v_i) + w_{place} \times sim_{place}(q, v_i),$$

where w_{object}, w_{action} and w_{place} are hyper-parameters for object, action and place. $sim_{object}(q, v_i)$ is the word2vec similarity between the query sentence q and the concepts of the video shot v_i in object aspect, and it is likewise in the $sim_{action}(q, v_i)$ and $sim_{place}(q, v_i)$. After giving similarity scores for all video shots, we can generate a ranked list of video shots in a descending order by their corresponding similarity scores. We test the model on the VBS 2019 development data, and use $w_{object} = 2$, $w_{action} = 4$ and $w_{place} = 1$ in our submission.

Type	Model name	Dataset	No. concept	F_M_C_D_ VIREO.19_1	F_M_C_D_ VIREO.19_2
Object/person	ResNet152 [6]	ImageNet Shuffle [7]	12988	x	x
Object/person	ResNet152	ImageNet [8]	1000	x	x
Object/person	FasterRCNN [9]	OpenImage V4 [10]	600		x
Object/person	ResNet152	TRECVID SIN Task [11]	346	x	x
Object/person	ResNet152	Research Collection [12]	497	x	x
Object/person	ResNet152	MSCOCO [13]	80	x	
Action	C3D [14]	Sport1M [15]	487	x	
Action	P3D [16]	Kinetics [17]	600	x	x
Action	P3D	ActivityNet [18]	200	x	
Place	ResNet152	MIT Places [19]	365	x	x

Table 1: Concept bank and training models used in concept-based approaches.

1.1.2 F_M_C_D_VIREO.19_2

In this run, we focus on the concept selection process. At first, we extract the uni-gram, bi-gram, and tri-gram from the query using an unsupervised model trained with *text8* corpus and embed all these n-grams and the original query by the *Universal Sentence Embedding* [3]. Similarly, we embed all the concept names using the same embedding model. For the concepts that belong to the ImageNet dataset, the concept names are generated by their WordNet synset [20].

In the concept selection step, for each item in the n-grams, we select a set of concepts by picking the nearest concepts in the embedding space by using a threshold r. Starting from the nearest concept, we *incrementally add more concepts* [4] from this set and we stop this process when the new added concept drifts out. We apply the same process with the embedding of the whole query sentence. At the end, we use the intersection of two sets of concepts (selected by n-grams and selected by the query sentence) to rank the video segments. If the intersection of these two sets is an empty set, we use the set of concepts selected by the query sentence embedding.

The concept bank used in this run is slightly different compare to the run *F_ M_ C_ D_ VIREO.19_ 1* (see Table 1) with more object detectors and less activities detectors. In addition, we use the same system but skip extracting the n-grams for concept selection to generate the result of the novelty run *F_ M_ N_ D_ VIREO.19_ 5.* The novelty run is our best automatic run and rank second among all the novelty runs in term of novel shots discovered.

1.2 Results analysis

Our benchmark results show that (1) the user intervention in query formulation boosts the performance of video search; (2) the concept-based search systems fail if the query is out of the concept bank's vocabulary.

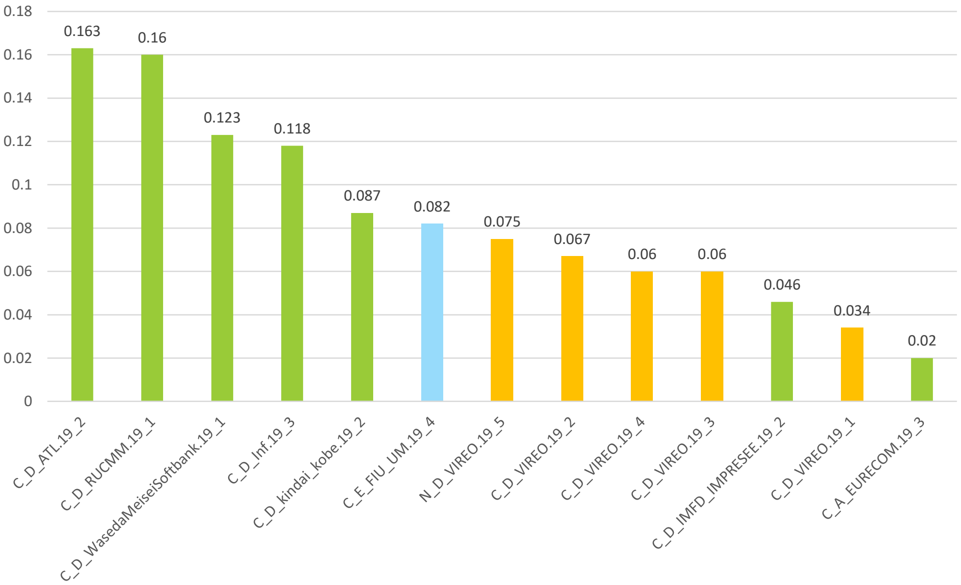

Figure 1: Performance comparison of official auto runs.

In Figure 1, we plot the best automatic runs from all the participants and all the automatic runs from VIREO team. An observation is that the combination of multiple runs does not improve but even drop down the performance if one of these runs is not good. This is the case of run *F_ M_ C_ D_ VIREO.19_ 3* and *F_ M_ C_ D_ VIREO.19_ 4*, where the performances are lower than the single run *F_ M_ C_ D_ VIREO.19_ 2*.

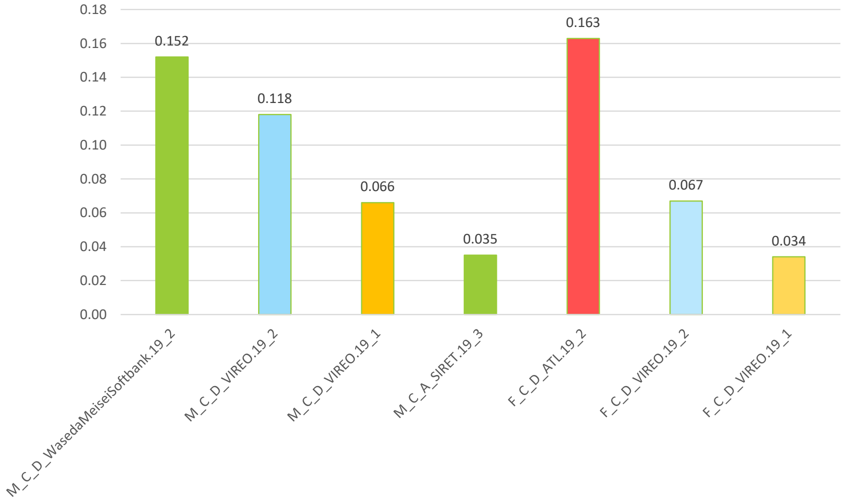

Figure 2: Performance comparison of official auto and manual runs.

Figure 2 presents the performance comparison between the auto and manual runs. For our concept-based searching systems, the manual runs are done by manually selecting the good concepts from the generated visual concepts. It is observed that the human intervention into the concept-based search can approximately double the performance of automatic runs (from 0.034 to 0.066 in the run *M_ M_ C_ D_ VIREO.19_ 1*, and from 0.067 to 0.118 in the run *M_ M_ C_ D_ VIREO.19_ 2*). The advantage of concept-based search is that the result is interpretable, where an end user can manipulate the concept list to make sense of a query. This is not possible for feature embedding learning however.

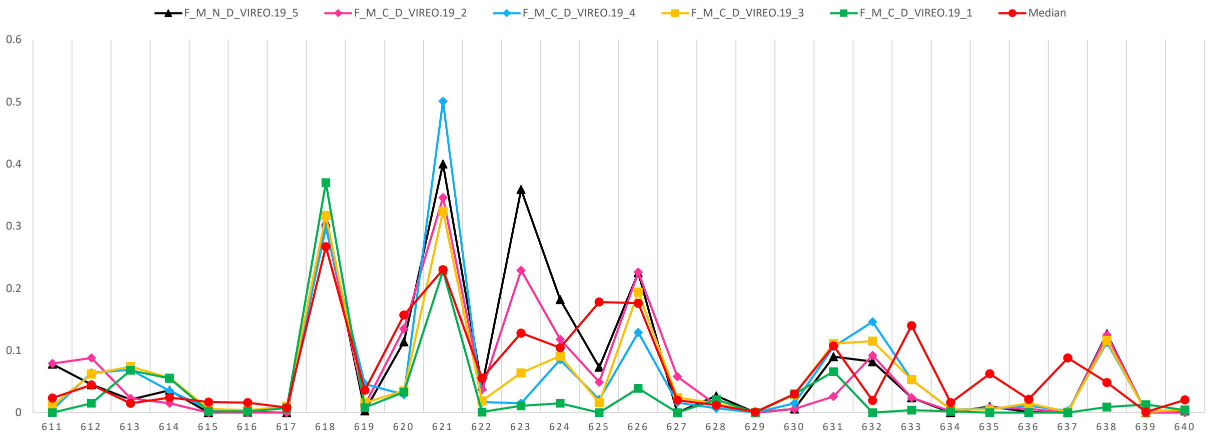

Figure 3: Performance comparison between automatic runs of VIREO team. The red circles show the median *xinfAP* of all the submissions.

In Figure 3, we plot the performance of our automatic runs in details and compare with the median performance of all submitted runs. In most cases, our runs perform better than the median *xinfAP* but they do not achieve the median *xinfAP* in few cases. In some cases, the concept-based search system fails because the critical concepts are out of the concept bank (out of vocabulary), such as "bald man" in query 635, "shirtless man" in query 637. In the query 611, the concept-based system confuses between "an unmanned aerial vehicle" and "a male honey bee" while trying to match a concept with "drone" and lead to different results. Also, our systems fail if the query contains detail properties of the object/person, such as "red dress" in query 616, "red hat or cap" in query 640. Those are shortcomings of our concept-based search systems that we are going to revise.

Acknowledgment

The work described in this paper was supported by a grant from the Research Grants Council of the Hong Kong SAR, China (Reference No.: CityU 11250716), and a grant from the PROCORE-France/Hong Kong Joint Research Scheme sponsored by the Research Grants Council of Hong Kong and the Consulate General of France in Hong Kong (Reference No.: F-CityU104/17).

References

[1] G. Awad, A. Butt, K. Curtis, Y. Lee, J. Fiscus, A. Godil, A. Delgado, A. F. Smeaton, Y. Graham, W. Kraaij, and G. Quénot, “Trecvid 2019: An evaluation campaign to benchmark video activity detection, video captioning and matching, and video search & retrieval,” in *Proceedings of TRECVID 2019*. NIST, USA, 2019.

[2] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, “The Stanford CoreNLP natural language processing toolkit,” in *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, Jun. 2014, pp. 55–60. [Online]. Available: https://www.aclweb.org/anthology/P14-5010

[3] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, “Universal sentence encoder,” *arXiv preprint arXiv:1803.11175*, 2018.

[4] M. H. T. D. Boer, Y.-J. Lu, H. Zhang, K. Schutte, C.-W. Ngo, and W. Kraaij, "Semantic reasoning in zero example video event retrieval," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 13, no. 4, pp. 60:1–60:17, Oct. 2017. [Online]. Available: http://doi.acm.org/10.1145/3131288

[5] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015.

[7] P. Mettes, D. C. Koelma, and C. G. Snoek, "The imagenet shuffle: Reorganized pre-training for video event detection," in *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*, 2016, pp. 175–182.

[8] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2014.

[9] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[10] A. Kuznetsova, H. Rom, N. Alldrin, J. R. R. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, T. Duerig, and V. Ferrari, "The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale," *CoRR*, vol. abs/1811.00982, 2018. [Online]. Available: http://arxiv.org/abs/1811.00982

[11] W. Zhang, H. Zhang, T. Yao, Y. Lu, J. Chen, and C.-W. Ngo, "Vireo@trecvid 2014: instance search and semantic indexing," in *In NIST TRECVID Workshop*, 2014.

[12] S. Strassel, A. Morris, J. Fiscus, C. Caruso, H. Lee, P. Over, J. Fiumara, B. Shaw, B. Antonishek, and M. Michel, "Creating havic: Heterogeneous audio visual internet collection," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12*, 2012.

[13] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 740–755.

[14] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4489–4497.

[15] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[16] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," *CoRR*, vol. abs/1711.10305, 2017. [Online]. Available: http://arxiv.org/abs/1711.10305

[17] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," *CoRR*, vol. abs/1705.06950, 2017. [Online]. Available: http://arxiv.org/abs/1705.06950

[18] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, "Activitynet: A large-scale video benchmark for human activity understanding," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015, pp. 961–970.

[19] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[20] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International journal of lexicography*, vol. 3, no. 4, pp. 235–244, 1990.