EURECOM at TRECVID 2017: The Adhoc Video Search

Danny Francis, Bernard Merialdo, Benoit Huet
Data Science Department, EURECOM, Sophia Antipolis, France
firstname.lastname@eurecom.fr

***Abstract*—This paper describes the submissions of the EURECOM team to the TRECVID 2017 AVS task. Our approach is to project both the text topic and the visual keyframes in the same vector space, corresponding to a word embedding. We compare and combine several word embeddings, also using specific keyword weightings.**

I. INTRODUCTION

EURECOM participated to the TRECVID 2017 Adhoc Video Search (AVS) [1]. The approach used in the AVS task was an evolution of the one used in AVS 2016 [2]. The main modifications were the use of the features vectors provided by CERTH [3], the combination of various word embeddings, the weighting of keywords based on visual context, and the interpolation of several combinations.

The AVS task requires to link the textual and visual contents. A topic is expressed as a sentence, and the task it to retrieve the shots in the test database which correspond to this topic. Four runs can be submitted, each run being a ranked list of at most 1,000 shots for each of the 30 test topics. Evaluation is performed using the usual Mean Inferred Average Precision measure.

For this task, the video collection is the Internet Archive IACC. The development data contains the IACC.1 and IACC.2 parts, which were processed in the previous SIN tasks. The test data is the new IACC.3 part, which was released for the first time this year for the AVS task. The development data comes with spare annotations of 310 concepts, which have been done collaboratively during the previous SIN tasks. The development data represents 1,400 hours of videos, about 1 million shots, and test data represents 600 hours of video, about 300,000 shots.

As examples of possible topics, the 48 queries of the 2008 task were provided. Also, the 30 topics used in 2016 are available, as well as the relevance judgments that were performed during the 2016 evaluation. It has to be noted that these relevance judgments are only a partial annotation of the video database, as only the shots necessary to estimate the inferred MAP for the submitted runs were evaluated.

II. DESCRIPTION OF OUR APPROACH

The AVS task requires to build models that link textual and visual data. Our approach is to project these two modalities in the same vector space, then use a simple Euclidean distance. In 2017, we use various word embedding vector spaces for this common vector space. Then we need mechanisms to project both the text topics and the visual keyframes in this vector space.

- the text topic is a sequence of words that have each a representation in the word embedding space. We build the vector representation of the whole topic as the average of these word representations. Eventually, the average will be weighted, with weights that have to be defined with respect to the importance of each keyword for the visual representation of the sentence.
- the test keyframes are in a completely different space. We use existing image analysis models to build a text based representation of each keyframe. In 2017, we use the following models:
    - the NeuralTalk [4] package generates sentences describing the visual content of images. It is applied to each keyframe, then the corresponding sentence is projected in the word embedding space using the same mechanism as for the text topics.

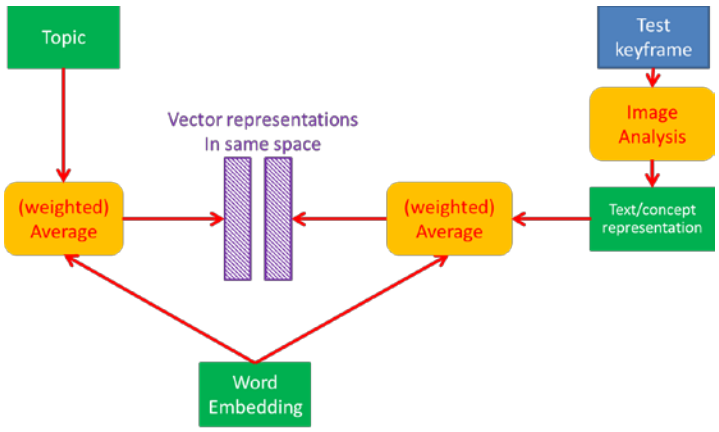
- the VGG Deep Networks [5] have been trained on part of the ImageNet database and can analyze an image to provide scores for 1,000 predefined concepts. Each concept is describe by a keyword or short description, therefore can be projected in the word embedding space. The average of these projections, weighted by the scores provided by the network, is the representation of the image in the word embedding space.
- the ImageNet Shuffle [6] provides similar classifiers, but they trained on a larger share of the ImageNet database and analyze images to produce scores for up to 13,000 concepts. The construction of the image representation is done in the same way as the previous method.
- In addition to the previous representation, we use the feature vectors provided by CERTH [3] for the test collection IACC3 using three different techniques:
  - Scores for the 1,000 ImageNet concept obtained with a combination of 5 pretrained classifiers,
  - Score for the TRECVID 345 Semantic Indexing (SIN) concepts, obtained by fusing the scores of fine-tuned neural classifiers,
  - Score for the TRECVID 345 Semantic Indexing (SIN) concepts, obtained by training a SVM classifier for each concept.

  These features vectors have been made available by CERTH to all participants in the Trecvid AVS task.

For the word embedding, we compared a number of existing systems, and after comparison on the TV16 concepts, we retained the following three:

- Parallel Document Context (PDC) [7], use syntagmatic and paradigmatic relations in the text to construct the model. We use the model trained on Wikipedia, with a dimension of 300.
- The Meta-Embeddings [8] which are built by combining other pre-existing embeddings.
- The LexVec embedding model [9] that uses specific matrix factorization and sampling techniques to build improved word representations.

The global strategy can be represented by the following diagram:



This diagram illustrates how topics and keyframes are projected in the same vector space, corresponding to the chosen word embedding. Several such diagrams are activated, using various word embeddings, image analysis models, and weights for the average, so that for given a topic and keyframe, we have several possible vectors representations, which in turn provide several scores to estimate the proximity between the topic and the keyframe.

After our participation in TV16 AVS we noticed that simply averaging word embeddings to make a sentence embedding was less efficient when sentences contained visually ambiguous words such as articles, pronouns or general concepts. As a result of that observation we decided to derive weights associated to their visual explicitness. For that purpose we used the MSCOCO database [10], containing 40k images with 5 sentence labels each. We computed 1k ImageNet scores for each image using a VGG Deep Network thus obtaining 1k dimensional vectors. If I is an image, let V_I denote its corresponding vector of ImageNet scores. Let w be a word and let S_w be the set of all image-sentence couples (I, s) in MSCOCO with s containing w. Eventually let $\bar{V}$ be the average vector of all images in MSCOCO. Then we derived word scores according to the following formula:

$$\text{score}(w) = \left\| \bar{V} - \frac{1}{|S_w|} \sum_{(I,s) \in S_w} V_I \right\|_2 .$$

We found that these scores worked like a visual tf-idf weighting: they were high both when words were uncommon and when they were explicitly designating a visual element of an image.

III. DESCRIPTION OF THE AVS RUNS

We tried a number of combinations, and use the TV16 AVS data to evaluate their performance. Based on these results, we submitted four runs to TV17 AVS. We describe these four runs in the following sections. The runs are numbered from 1 to 4, with the expected best runs having the lower numbers. Therefore, we describe the runs by decreasing number

A. *RUN 4 "Single"*

This run is based on the best combination of methods that we measured on TV16. It is based on the CERTH ImageNet scores for representing the visual content, and the PDC 300 word embedding. The distance in the word embedding space is the usual Euclidean distance, and the vectors are L2-normalized.

B. *RUN 3 "Regular Merge"*

In this run, we consider 18 different combinations, and we optimize the weight of the linear combination of scores to provide a maximal performance on TV 16. The combinations use all 6 possible visual representations described in the previous chapter, combined with the 3 possible embeddings described previously. The 18 weights are optimized by the leave-one-out technique on the TV16 data. Again, the distance is the Euclidean distance, and all vectors are L2-normalized.

C. *RUN 2 "Weighted Merge"*

In this run, we change the construction of the topic vector representation by using the visual weights that we have previously defined. The other components of the processing are the same as in RUN 3.

D. *RUN 1 "All Merge"*

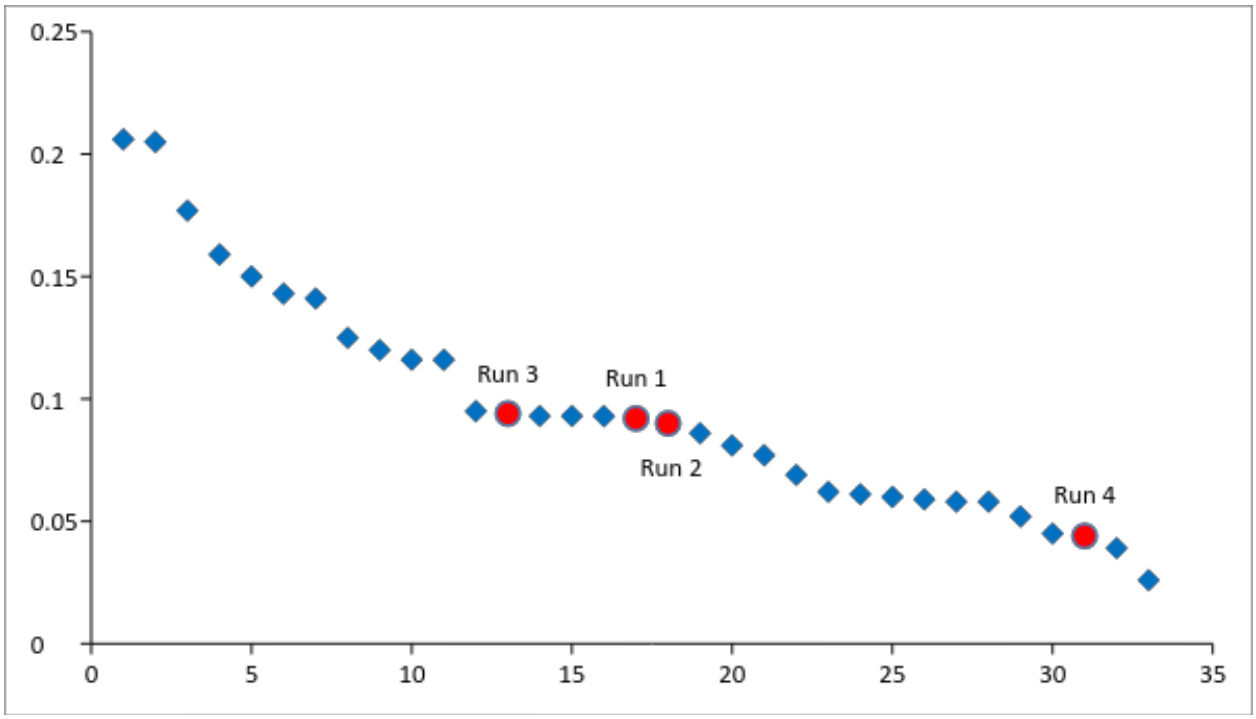
This run is quite similar to RUN 3, but we also consider the un-normalized vectors. This leads to 36 combinations of visual vectors, word embedding and normalization type. The weights for the linear combination of these 36 scores are optimized with the same strategy as RUN 3.

IV. AVS RUNS EVALUATIONS

The result (MAP) obtained by our four runs are the following:

TEAM	RUN	MAP
EURECOM	3	0,094
EURECOM	1	0,092
EURECOM	2	0,090
EURECOM	4	0,044

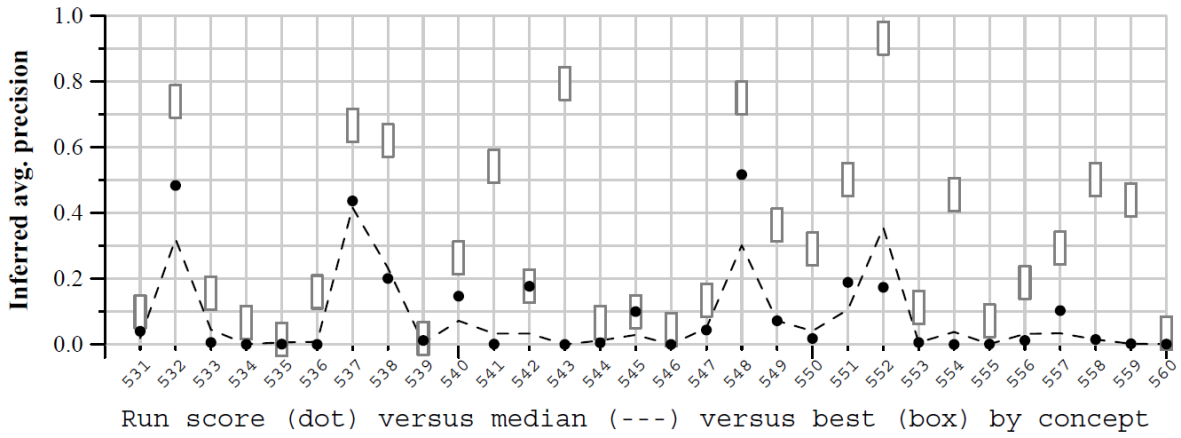
The following graph shows how they are located within the full set of 33 (Fully Automatic) submissions from all participants :



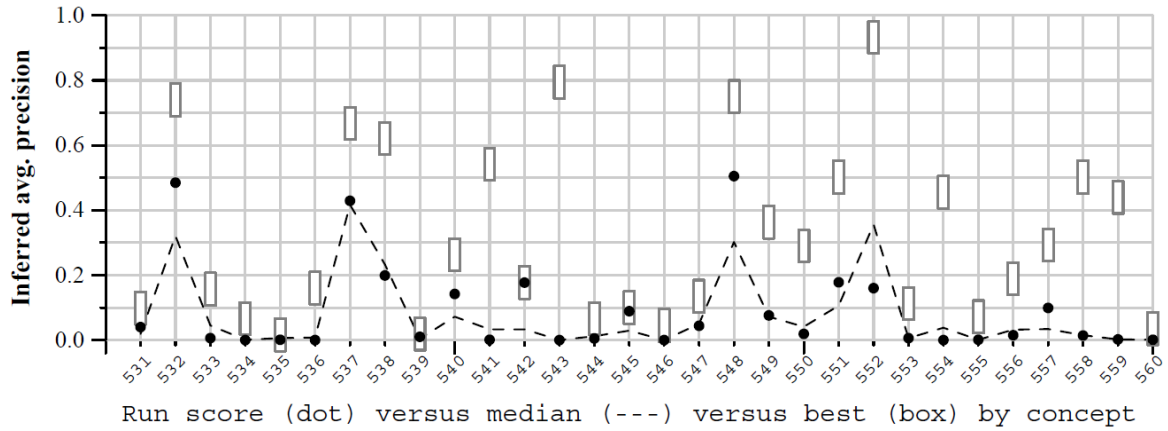
We can observe that our best run is RUN 3, which is based on the linear combination of 16 scores obtained with the different visual representations and word embeddings. Normalization plays an important role, as the extra scores without normalization that are included in RUN 1 actually degrade the performance. We were disappointed that the visual weighting of keywords used in RUN 2 did not allow to improve over a simple average. One reason maybe that the data used to define the visual weights is not sufficiently relevant to the topics used in TV17. As expected, RUN 4 which uses a single combination has a lower performance, but the gap with the results obtained by the other runs show the tremendous effect of the combination of techniques.

The detailed performances of our runs on each topic are shown in the following figures:

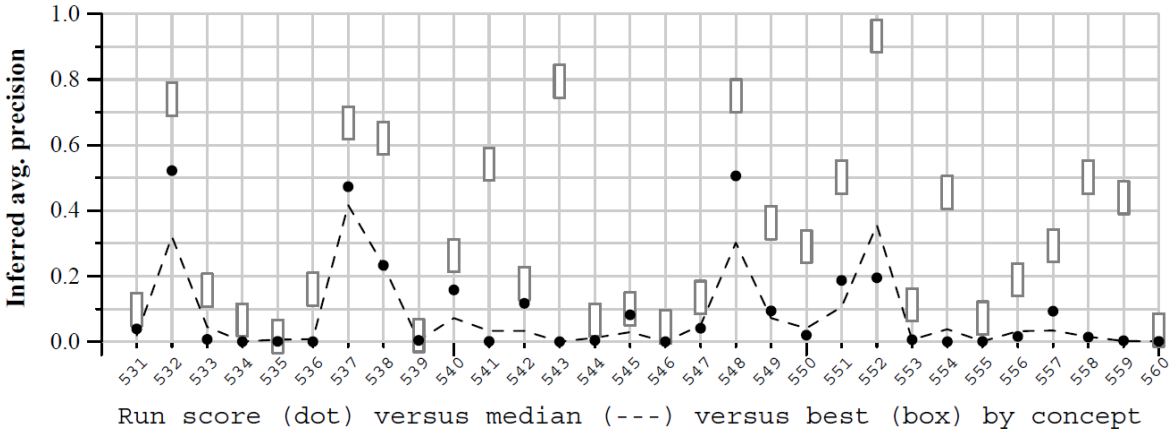
RUN 1:



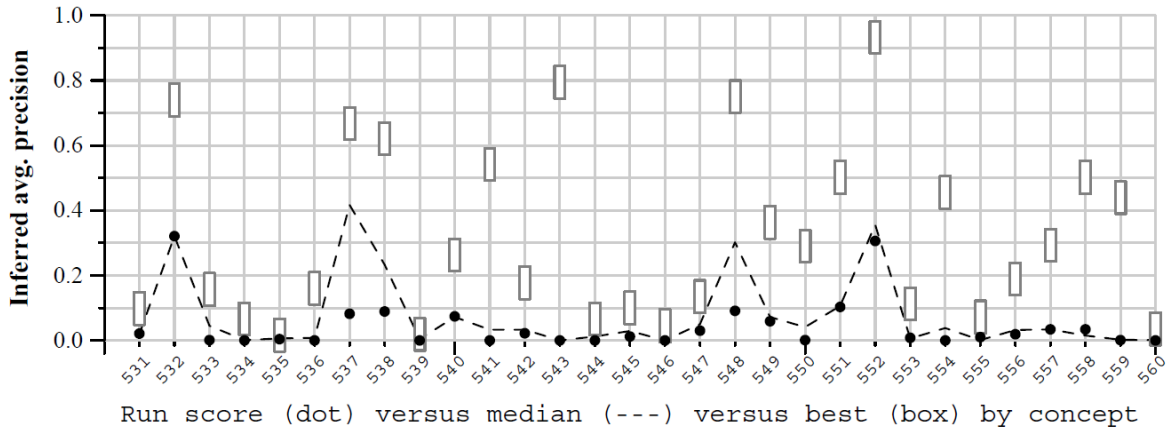
RUN 2:



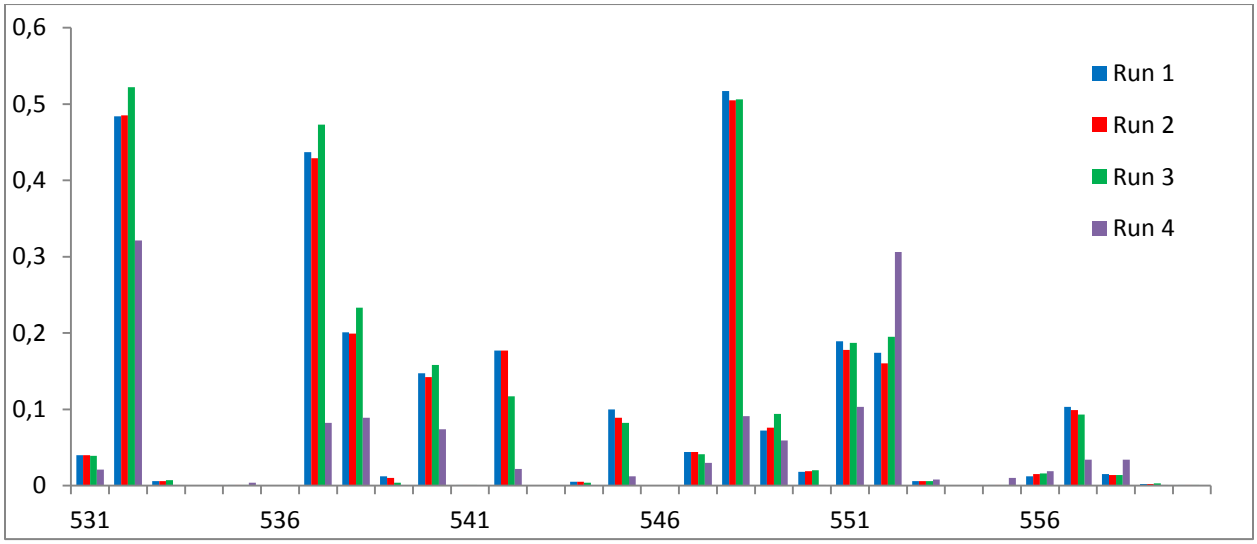
RUN 3:



RUN 4:



The following figure shows the comparative results of our 4 runs per topic:



These graphs show that there is a great discrepancy in the performance depending on the topic. For some topics, the performance are quite reasonable, for others, the performance is very close to zero. We will later investigate the relation between the semantic meaning of the topics and the observed performance, to try to find any useful correlation.

ACKNOWLEDGEMENT

Part of this work was done within the scope of the ANR GAFES project.

REFERENCES

[1] George Awad, Asad Butt, Jonathan Fiscus, David Joy, Andrew Delgado, Martial Michel, Alan F. Smeaton, Yvette Graham, Wessel Kraaij, Georges Quénot, Maria Eskevich, Roeland Ordelman, Gareth J. F. Jones, Benoit Huet, TRECVID 2017: Evaluating Ad-hoc and Instance Video Search, Events Detection, Video Captioning and Hyperlinking, Proceedings of TRECVID 2017, 2017, NIST, USA

[2] Merialdo Bernard, Pidou Paul, Eskevich Maria, Huet Benoit, EURECOM at TRECVID 2016: The Adhoc Video Search and Video Hyperlinking tasks, TRECVID 2016, 20th International Workshop on Video Retrieval Evaluation, 14-16 Novembre 2016, Gaithersburg, Ma, USA

[3] Markatopoulou, Foteini, & Mezaris, Vasileios. (2017). Concept detection scores for the IACC.3 dataset (TRECVID AVS Task) [Data set]. Zenodo. http://doi.org/10.5281/zenodo.292994

[4] Andrej Karpathy, Li Fei-Fei, Deep Visual-Semantic Alignments for Generating Image Descriptions, CVPR 2015

[5] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556, 2014.

[6] Pascal Mettes, Dennis C. Koelma, and Cees G.M. Snoek. 2016. The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection. In Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16). ACM, New York, NY, USA, 175-182.

[7] Fei Sun, Jiafeng Guo, Yanyan Lan, Jun Xu, Xueqi Cheng, Learning Word Representations by Jointly Modeling Syntagmatic and Paradigmatic Relations, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing , 2015, Beijing, China

[8] Wenpeng Yin, Hinrich Schütze, Learning Word Meta-Embeddings, Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany.

[9] Salle, Alexandre, Marco Idiart, Aline Villavicencio, Matrix Factorization using Window Sampling and Negative Sampling for Improved Word Representations. The 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany.

[10] Lin TY. et al. (2014) Microsoft COCO: Common Objects in Context. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer