EURECOM at TRECVID 2016: The Adhoc Video Search and Video Hyperlinking Tasks

Bernard Merialdo[1], Paul Pidou[1], Maria Eskevich[2], Benoit Huet[1]
[1]Data Science Department, EURECOM, Sophia Antipolis, France
[1]firstname.lastname@eurecom.fr
[2]Radboud University, The Netherlands
[2]m.eskevich@let.ru.nl

***Abstract*—This paper describes the submissions of the EURECOM team to the TRECVID 2016 AVS and LNK tasks.**

I. INTRODUCTION

EURECOM participated to the TRECVID 2016 Adhoc Video Search (AVS) and Video Hyperlinking (LNK) tasks [1]. We used to participate in the Semantic Indexing (SIN) task, but this task was discontinued. The AVS is a new task (except for a trial in TRECVID 2008), which required to design and implement new mechanisms compared to our previous works. In case of the LNK task, we have followed the approach that was implemented in our 2015 submission [14], extending it with diverse normalization schemes.

II. AVS TASK FRAMEWORK

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.

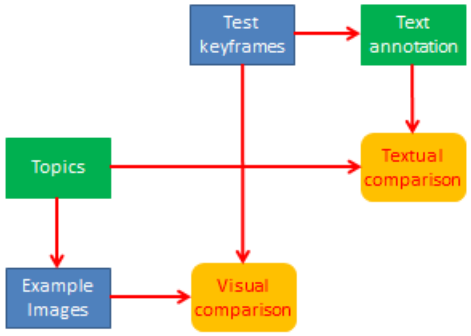
As examples of possible topics, the 48 queries of the 2008 task were provided. However, no other information was available, in particular, no example of successful images for these topics were available. Yet, different types of submissions were allowed, one of them including the use of automatic tools that are available on the internet, provided they do not include human intervention on the data submitted.

Since the AVS task requires to build models that link textual and visual data, we explored two possible strategies:

- from the text topic, interrogate web image search engines to collect examples of relevant pictures, then use these pictures to build a visual model, which in turn will select the best keyframes in the test database.

- from the test keyframes, automatically generate a text description, and then match this text description with the topic.

These two strategies are illustrated in the following figure :



In order to implement these strategies, we used the following tools and services, which are freely available from the internet:

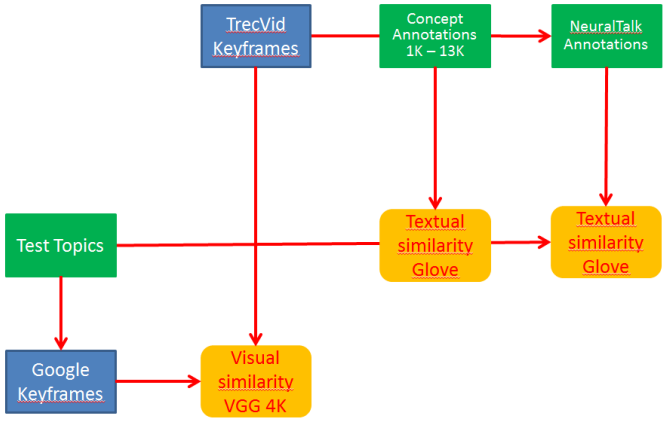
- to get example images for a topic, we used the Google ImageSearch engine [2]. This search engines allows to enter a text query and returns a list of corresponding images. The exact mechanism to retrieve those images is not published, however it is likely to be largely based on the textual context of the pages where these images appear. Although a number of other image search services are available, we limited ourselves to this only one by lack of time. For each topic, we kept only the first 100 images returned, as more and more irrelevant images occur when we go deeper in the result list.
- to get a text description from an image, we used several tools:
  - the VGG Deep Networks [3], which have been trained on part of the ImageNet database and can analyze an image to provide scores for 1,000 predefined concepts,
  - the ImageNet Shuffle [4], which provides classifiers trained on a larger share of the ImageNet database, and analyze images to produce scores for up to 13,000 concepts
  - the NeuralTalk [5] package, which generates sentences describing the visual content of images.
- to compare visual contents, we compute a visual feature vector for an image by apply the VGG Deep Network to each image and extract the outputs of the one-before-last and two-before-last layers, to build visual vectors. The similarity between visual vectors is computed as the usual scalar product, sometimes with normalization.
- to compare textual content, we use the GloVe vector representations of words [6], to build a textual vector from either the topic description, the concept name or the descriptive sentence. The similarity between textual vectors is again computed as the usual scalar product.

Many combinations of these modules are possible, as well as different values of the parameters involved. In order to choose the combinations to be used in the final runs, we performed a number of experiments on the development collection. We ran several systems using the 48 development topics, and applied them on the development videos. Then, we manually annotated the 10 best keyframes returned for each system and each topic. This gave us some indications of which system would have the greater performance. We observed that the performance of very different approaches was very depending on the topic, so in the final runs, we also chose to provide a selection of the different combinations that we tried.

III. DESCRIPTION OF THE AVS RUNS

A. *Generic Architecture*

The following figure illustrate the generic architecture that we have put in place, corresponding modules. The green modules represent text-based information, the blue modules contain visual information, the yellow modules represent similarity computations. We tried various combinations to define the four runs that we submitted to the final evaluation.



All our runs are of the "Fully Automatic" category, since no manual processing was done at any stage, and with the "D" training type, as we are using tools which were trained on data external to TRECVID.

B. *RUN 1 "GoogleSearch + VGG 4K"*

For each of the topic, we performed a search using the Google Image engine, and retained the first 100 pictures of the ranked list. To each image, we applied the VGG Deep network, and kept the one-before-last layer as feature vector of dimension 4K. We applied the same visual processing to each of the TRECVID keyframes in the test collection, and ranked them according to a Nearest Neighbor distance from the Google images.

C. *RUN 2 "ImageShuffle + Glove300"*

We used the ImageShuffle system to obtain scores for 13,000 concepts, which we used as feature vectors for each TRECVID keyframe. We used these scores as weights to compute a semantic vector of dimension 300 by a linear combination of the 13,000 Glove vectors corresponding to the concepts. For each topic, we constructed a semantic vector of dimension 300 by averaging the Glove vectors of the words appearing in the topic. Then we used the cosine similarity to find the images whose semantic vectors were most similar to the topics.

D. *RUN 3 "NeuralTalk + Glove300"*

We used the NeuralTalk system to generate text descriptions for each of the TRECVID keyframes. Then, we built a semantic vector of dimension 300 by averaging the Glove vectors of the words appearing in these descriptions. We did the same for the test topics. Finally, we used again the cosine similarity to find the images whose semantic vectors were most similar to the topics.

E. *RUN 4 "Global Average"*

During the development phase, we experimented with a number of combinations of the modules that we have described, using different dimensions, different projections, different layers,

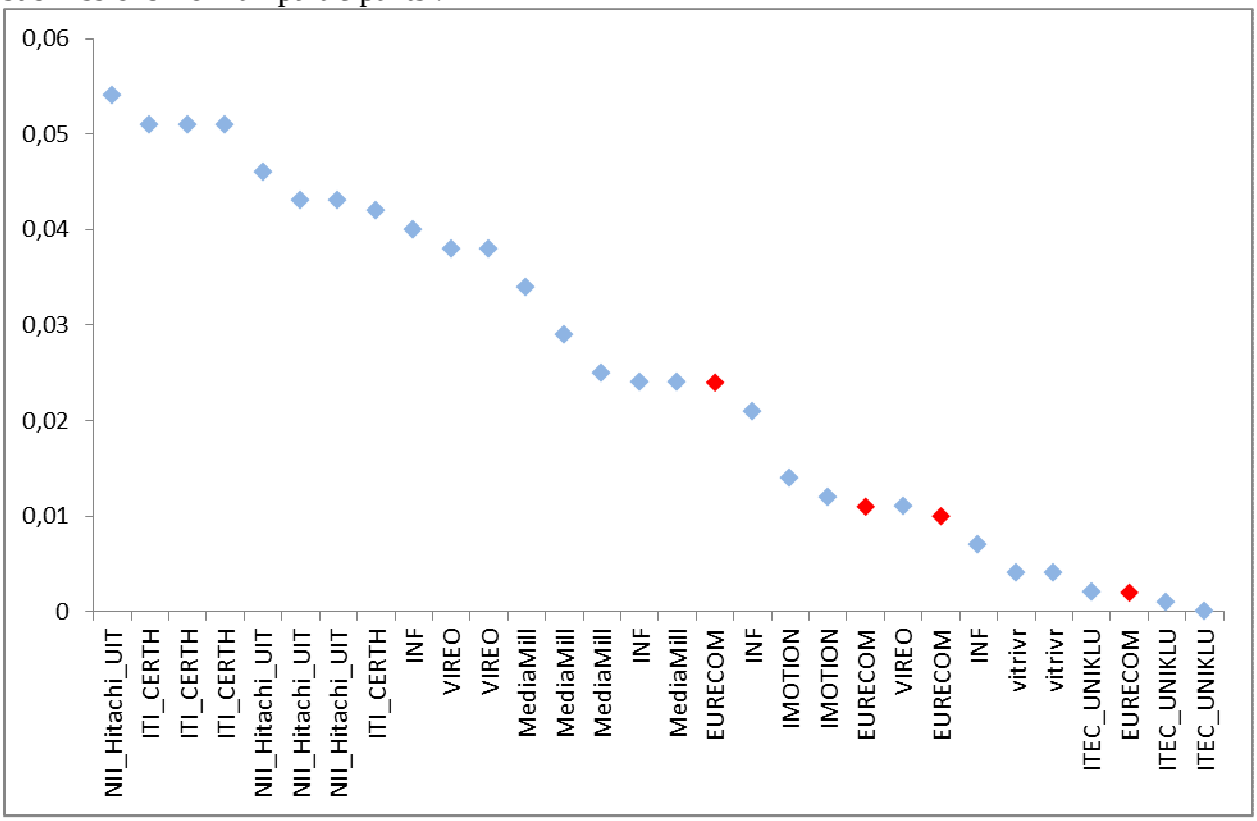
different similarity measures. We evaluated these combinations with a minimal annotation on the development collections, by pooling the 10 best pictures for each of the training topics. This gave us an indication of which combinations could be the most efficient, and helped us in the selection of the combinations for the final runs to be submitted. As we noticed that different combinations had very different performances of different topics, we tried to get the best of all combinations by averaging the results of 32 combinations that we had found to be of reasonable performance. As the similarity scores are not always comparable between different combinations, we introduced for each combination an artificial score computed as the inverse rank of each image in the result list. The average of these 32 inverse ranks is the final score for this run.

IV. AVS RUNS EVALUATIONS

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

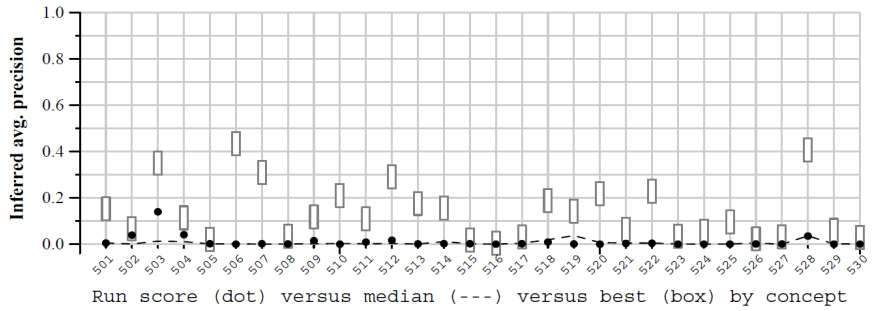
TEAM	RUN	MAP
EURECOM	2	0,024
EURECOM	1	0,011
EURECOM	4	0,01
EURECOM	3	0,002

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

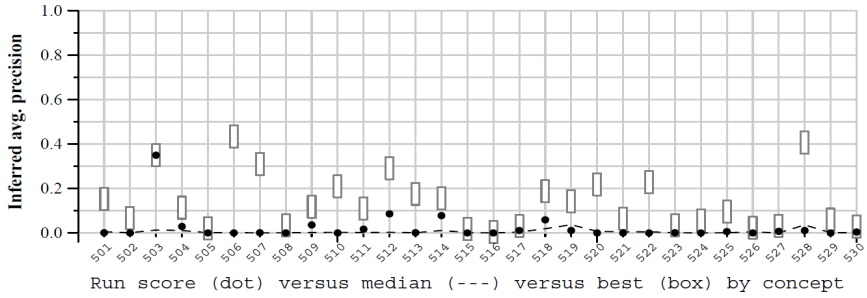


We can observe that our best run is RUN2, which is based on the ImageShuffle system, and has obtained a performance quite similar to the MediaMill team (which has developed ImageShuffle). The runs using Google Search or the full average have surprisingly very similar performance. Run3, based on NeuralTalk, performed quite poorly, probably because of the mismatch between the test topics and the type of annotations on which NeuralTalk was trained.

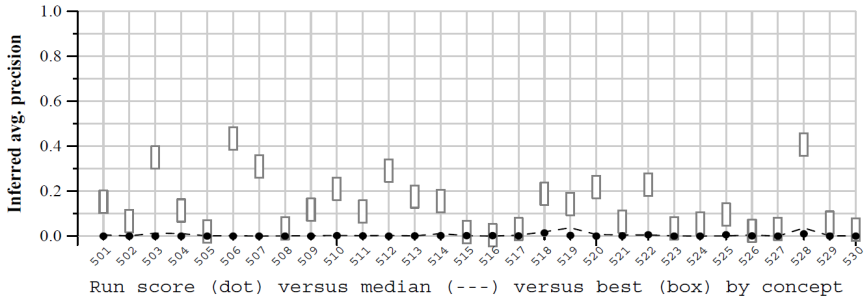
The detailed performances of our runs on each topic are shown in the following figures:
RUN1:



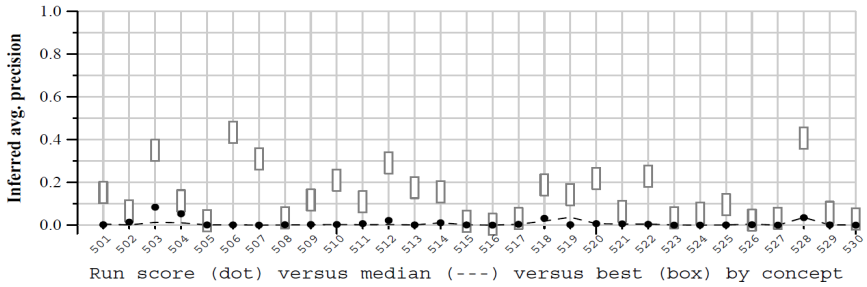
RUN2



RUN3



RUN4



These show that the good performance of RUN2 is probably due to its good score on topic 503, which is “Find shots of a person playing drums indoors”, although we have no specific interpretation for this result.

V. LNK TASK FRAMEWORK

Video Hyperlinking task in 2016 kept the main framework of video-to-video search between anchor and target segments, with 2 main differences: 1) the dataset was changed (instead of the professionally created and curated broadcast content provided by BBC, a collection of semi-professional user-generated videos (crawled from the blip.tv website) was used [1]; 2) the anchor video segments were chosen to reflect the uploader’s intent and to be of truly multimodal nature, i.e. a combination of both audio and visual streams is crucial for the anchor understanding, processing, and target selection [2].

The video collection consisted of 14,838 items that were taken from the Blip10000 dataset [17]. Released shot segmentation and corresponding keyframes were extracted at the stage of original collection creation [24]. As part of the 2016 collection release, state-of-the-art automatic speech recognition (ASR) transcripts [23] and extracted visual features were made available to the task participants. These visual concepts have been obtained running the BLVC CaffeNet implementation of the AlexNet [18], which was trained by Jeff Donahue (@jeffdonahue) with minor variation from the version described in [18]. For each shot of the video collection a key-frame is extracted and fed to the deep network for classification over the 1000 ImageNet Concepts. The top five concepts are provided for each key-frame along with their scores.

As participants, we have also extracted the visual features using the same principle (one key-frame per video shot) with the GoogleNet deep network architecture [22] which was shown to provide better accuracy on the ImageNet challenge.

VI. LNK SYSTEM SET UP

A. *Generic Architecture*

As the task has a new dataset and slightly different anchor creation strategy this year, we could not directly compare the results in case of implementation of the same methods as developed in 2015. However, as we are interested in tracing the patterns of video-to-video search performance across datasets and variety of users interests, we did follow the similar generic architecture in our approach, adding scores normalization and testing other visual features than the ones provided by organizers.

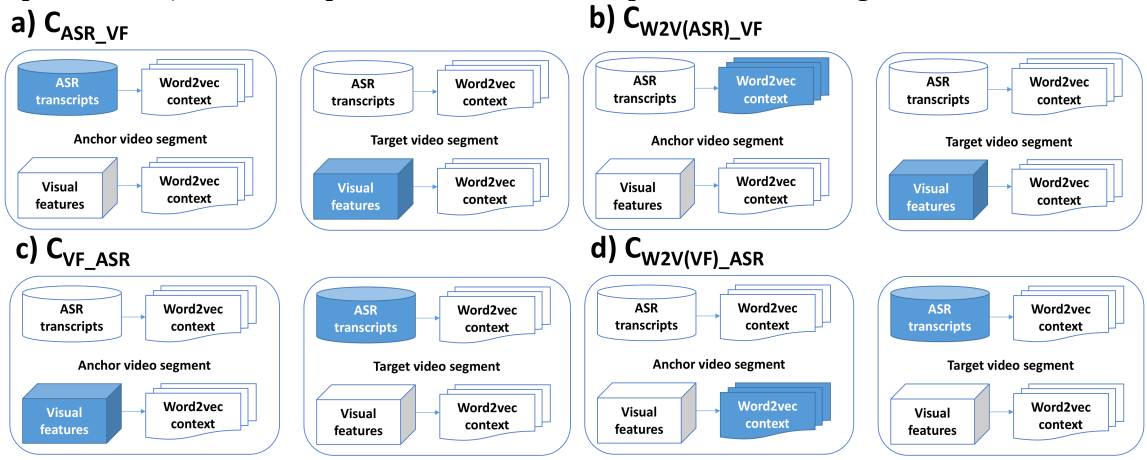
The system set up is based on the open source Terrier Information Retrieval tool that we use for initial indexing and retrieval, while the ranked list is further readjusted based on the visual features extracted for both videos and their connection to the audio content, and their contexts that are defined using word2vec terms proximity.

B. *Hyperlinking in Steps*

First, we split all the videos in the collection into fixed length segments of 120 seconds with a 30 seconds overlap step. We use these sharp time boundaries for all the features calculations, while we store the information about the start of the first word after a pause longer than 0.5 seconds or a first switch of speakers as the potential jump-in point for each segment of the segments, and those are used as starting point in the final submission runs, as in [19].

Second, we use the open-source Terrier 4.0. Information Retrieval platform[1] [20] with a standard language modeling implementation [21], with default lamda value equal to 0.15, for indexing and retrieval.

Thirdly, we calculate a set of additional confidence scores that reflect the connection between the verbal-visual content of both anchor and target, for all the retrieved top 1000 segments for each of the 90 anchors. These confidence scores consist of 4 components, as depicted in Figure below: a) Average confidence score of terms representing the visual concepts extracted for all the frames of the target video that overlap with the terms of the ASR transcript terms of the anchor video; b) Average confidence score of terms that represent the context of the anchor ASR transcript (word2vec top 100 items) and overlap with terms representing the visual concepts extracted for all the frames of the target video; c) Average confidence score of the terms representing the visual concepts extracted for all the frames of the anchor video that overlap with the ASR transcript terms of the target video; d) Average concept score of the terms that represent the context of the terms representing visual concepts extracted for the anchor video (word2vec top 100 items) that overlap with the ASR transcript terms of the target video.



Final ranking of the target segments for each anchor was based on the combination of the original Terrier ranking, that reflects verbal-verbal connection between anchor and target ASR transcripts, and the verbal-visual components. Each of the 5 confidence scores was normalized and then summed together with either equal weights for each component (0.2 for each component), runs 1 and 4; or with higher priority assigned to the components using the word2vec context (0.35 versus 0.1), runs 2 and 3.

Once these segments were re-ranked according to the combined confidence score, the overlapping ones were removed, and this resulted in four submitted runs.

For 2 runs (1 and 2) we have used GoogleNet visual concepts, and for the runs 3 and 4 we used the ones provided by the task organizers.

[1] http://www.terrier.org

VII. LNK RUNS EVALUATIONS

The results of the submissions were judged at top 5 ranks using a set of metrics: precision at rank 5 (P@5), and its variations (P@5_bin, P@5_tol), MAP, and its variations (MAP_bin, MAP_tol), MAiSP [25, 26].

Run ID	P@5	P@5_bin	P@5_tol	MAP	MAP_bin	MAP_tol	MAiSP
1	0.3356	0.3400	0.3244	**0.0761**	**0.0926**	**0.0632**	**0.1197**
2	0.3267	0.3244	0.3133	0.0707	0.0870	0.0561	0.1076
3	0.3422	0.3422	0.3222	0.0709	0.0864	0.0571	0.1096
4	**0.3511**	**0.3444**	**0.3333**	0.0759	0.0886	0.0618	0.1154

MAiSP metric takes into account the user experience of interacting with the multimedia content which is more time-consuming and demanding comparing to the scrolling of the textual retrieval results, and according to it all the submitted runs are in the top third of the results amongst all participants, run 1 being the second in the overall ranking.

The fact, that the value of the metric P@5 does not drastically change when the binning or window of tolerance are applied (P@5_bin, P@5_tol), implies that the result lists demonstrate a variety of retrieved target video segments without overlap or segments extracted from the neighborhood regions of the same videos.

REFERENCES

[1] George Awad, Jonathan Fiscus, Martial Michel, David Joy, Wessel Kraaij, Alan F. Smeaton, Georges Quénot, Maria Eskevich, Robin Aly, Roeland Ordelman, TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking, Proceedings of TRECVID 2016, 2016, NIST, USA

[2] https://www.google.fr/imghp?

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

[4] 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.

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

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning, GloVe: Global Vectors for Word Representation? Conference on Empirical Methods in Natural Language Processing, 2014

[7] E. E. Apostolidis, V. Mezaris, M. Sahuguet, B. Huet, B. Cervenkov´a, D. Stein, S. Eickeler, J. L. R. Garc´ıa, R. Troncy, and L. Pikora. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014, pages 1033–1036, 2014.

[8] S. Ayache and G. Quenot. Video Corpus Annotation using Active Learning. In European Conference on Information Retrieval (ECIR), pages 187–198, Glasgow, Scotland, mar 2008.

[9] M. Eskevich and B. Huet. EURECOM @ SAVA2015: Visual Features for Multimedia Search. In MediaEval 2015 Multimedia Benchmark Workshop, Wurzen, Germany, 2015.

[10] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093, 2014.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolu- tional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Asso- ciates, Inc., 2012.

[12] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representa- tions of words and phrases and their compositionality. In Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Sys- tems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States., pages 3111–3119, 2013.

[13] U. Niaz, B. Merialdo, and C. Tanase. EURECOM at TRECVID 2014: The semantic indexing task. In TRECVID 2014, 18th International Workshop on Video Retrieval Evaluation, 10-12 November 2014, Orlando, USA, Orlando, UNITED STATES, 11 2014.

[14] U. Niaz, B. Merialdo, C. Tanase, M. Eskevich, and B. Huet. EURECOM at TRECVID 2015: The semantic indexing and video hyperlinking tasks. In TRECVID 2015, 19th International Workshop on Video Retrieval Evaluation, 2015, National Institute of Standards and Technology, Gaithersburg, USA, Gaithersburg, UNITED STATES, 11 2015.

[15] B. Safadi, M. Sahuguet, and B. Huet. When Textual and Visual Information Join Forces for Multimedia Retrieval. In Proceedings of International Conference on Multimedia Retrieval (ICMR'14), pages 265:265–265:272, Glasgow, United Kingdom, 2014.

[16] M. Eskevich, M. Larson, R. Aly, S. Sabetghadam, G.J.F. Jones, R. Ordelman, and B. Huet. Multimodal Video-to-Video Linking: Turning to the Crowd for Insight and Evaluation. In Proceedings of 23rd International Conference on Multimedia Modeling, January 4-6, 2017, Reykjavik, Iceland.

[17] S. Schmiedeke, P. Xu, I. Ferrané, M. Eskevich, C. Kofler, M.A. Larson, Y. Estève, L. Lamel, G.J.F. Jones, and T. Sikora. Blip10000: a social video dataset containing SPUG content for tagging and retrieval. In Proceedings of ACM Multimedia Systems Conference 2013, pages 96-101, Oslo, Norway, 2013.

[18] A. Krizhevsky, I. Sutskever, and G.E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25, pages 1097 - 1105, 2012.

[19] M. Eskevich, G.J.F. Jones. Time-based segmentation and use of jump-in points in DCU search runs at the Search and Hyperlinking task at MediaEval 2013. In Proceedings of MediaEval 2013 Multimedia Benchmark Workshop, Barcelona, Spain, 2013.

[20] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006), 2006.

[21] D. Hiemstra. Using language models for information retrieval. PhD thesis, University of Twente, The Netherlands, 2001.

[22] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A., "Going deeper with convolutions." In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9. 2015.

[23] L. Lamel. Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data. In Proceedings of the Fifth International Conference Baltic HLT 2012, pages 1-8, Tartu, Estonia, 2012.

[24] P. Kelm, S. Schmiedeke, and T. Sikora. Feature-based video key frame extraction for low quality video sequences. In Proceedings of 10th Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), pages 25-28, London, United Kingdom, 2009.

[25] R. Aly, M. Eskevich, R. Ordelman, and G.J.F. Jones. Adapting Binary Information Retrieval Evaluation Metrics for Segment-based Retrieval Tasks. CoRR: abs/1312.1913. 2013. http://arxiv.org/abs/1312.1913.

[26] D. N. Racca, and Gareth J. F. Jones. Evaluating Search and Hyperlinking: An Example of the Design, Test, Refine Cycle for Metric Development. In Proceedings of the MediaEval 2015 Workshop, Wurzen, Germany, 2015.