The vitrivr System at TRECVID 2016: The Ad-Hoc Video Search Task

Claudiu Tănase Luca Rossetto Ivan Giangreco Heiko Schuldt
Department of Mathematics and Computer Science
University of Basel, Switzerland
{c.tanase | luca.rossetto | ivan.giangreco | heiko.schuldt}@unibas.ch

ABSTRACT

In this paper, we describe the details of our participation to the TRECVID Ad-Hoc Video Search (AVS) 2016 with the vitrivr system.

1. INTRODUCTION

This paper describes the details of our participation to the TREC Video Retrieval Evaluation 2016 [1] Ad-Hoc Video Search (AVS) Task with the vitrivr system. The AVS task considers an end-user looking for a video segment in a collection that has not been previously manually annotated. In this task, 30 queries were released by NIST, for which the system should return a ranked result list of at most 1000 shot IDs each. The test data set is composed of 4593 Internet Archive videos with a total duration of 600 hours (IACC.3).

In this paper, we present the vitrivr system. The vitrivr system [8] is an open-source video retrieval system[1] is powered by the ADAM$_{pro}$ database [2] and the Cineast retrieval engine [7].

This paper is structured as follows: Section 2 describes in detail the submitted runs and Section 3 presents a description of the system. In Section 4, we discuss the results of our runs. Finally, Section 5 concludes.

2. SUBMITTED RUNS OVERVIEW

The submitted runs have been named based on the priority ordering, so that 'Run 4' is the lowest priority run. Runs 4 and 3 are fully automated, runs 1 and 2 are manually assisted.

vitrivr_4 captions for the test keyframes were automatically generated using NeuralTalk2 and ranked based on cosine similarity to the query payload text in a 400 topic LSI text feature space trained on a recent Wikipedia text dump (see Figure 1).

vitrivr_3 4096-dimensionality feature vectors representing activations in the 7^{th} fully connected layer in a VGG16 neural network were extracted from IACC.3 keyframes, MSCOCO and Flickr30k images. Using the same LSI model as in Run4, we scored captions on the MSCOCO and Flickr30k and used these textual similarity scores as target values for random forest and linear support vector regressors. We fuse the test scores by summing. This run is illustrated in Figure 2.

vitrivr_2 4096-dimensionality feature vectors representing activations in the 7^{th} fully connected layer in a VGG16 neural network were extracted from IACC.3 keyframes, MSCOCO and Flickr30k images. 1957 training examples for the 30 queries were manually collected and RBF and chi-square kernel SVMs were trained on them. We fuse the test scores by summing. Details can be found in Figure 3.

vitrivr_1 Score fusion by simple summing scores of the runs 4, 3 and 2 (see Figure 4).

3. SYSTEM DESCRIPTION

The following sections describe individual system components involved in scoring the runs.

3.1 Extra training data

In order to establish a relationship between visual and semantic textual data, we made use of datasets containing images which are annotated with short textual descriptions created by humans. The two datasets used in this work were the MSCOCO [5] caption dataset consisting of roughly 120 thousand images with one caption per image and the Flickr30k [9, 6] dataset which contains 30 thousand images and has five captions per image.

In addition to the above mentioned datasets, we collected 1957 exemplary images which depict objects and scenes relevant to the queries resulting in 61 specialized training examples per query on average. The images were collected using Google image search. No restriction in size was imposed upon the images since all had to be rescaled in order to be processed. Details on the processing can be found in Section 3.2. For some queries, little to no images matching the query perfectly could be obtained. In these instances, images containing at least the most dominant aspects of the query were selected.

[1] https://vitrivr.org/

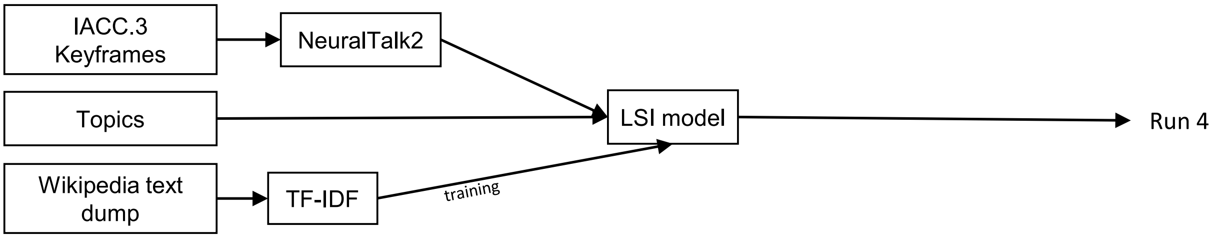

Figure 1: Flow diagram of Run 4

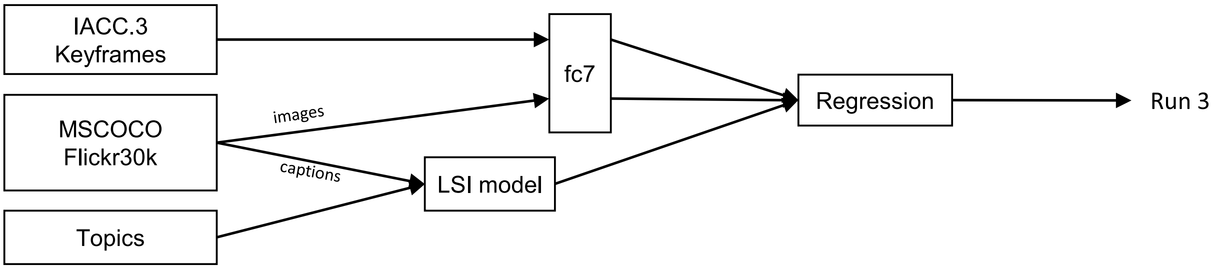

Figure 2: Flow diagram of Run 3

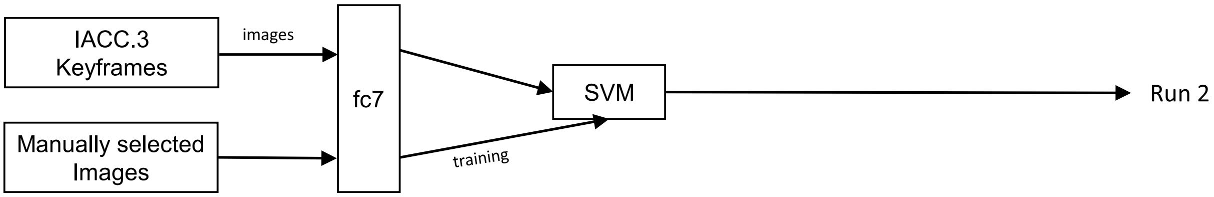

Figure 3: Flow diagram of Run 2

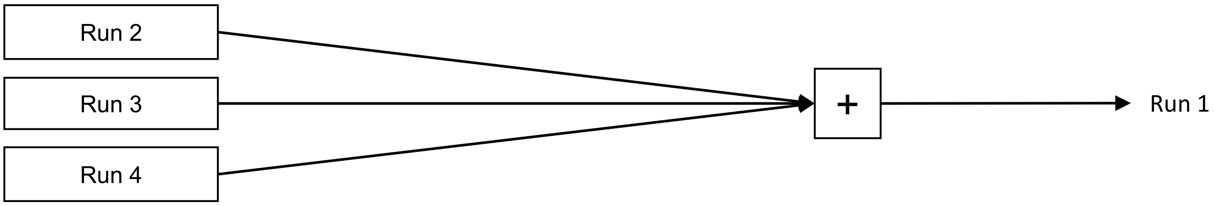

Figure 4: Flow diagram of Run 1

3.2 CNN feature extraction

To obtain features capable of describing the semantic content of an image, we use output of the 7^{th} fully connected layer of a neural network [4]. We used a pre-trained model based on the BVLC CaffeNet Model[2] but converted for a CPU-based DNN runtime[3]. The output of this layer is a 4096-dimensional sparse vector.

3.3 Automated captioning and retrieval

We used NeuralTalk 2 [3] for automatic image captioning. This was done using the CPU version of the pre-trained network provided by the authors of NeuralTalk[4]. Using this network, one caption was generated for every IACC.3 keyframe.

In order to estimate semantic similarity between text captions we build our own text retrieval module. We use as training corpus a recent 13GB text dump of Wikipedia[5] on which we perform simple tokenizing (without stemming). We build a dictionary of the 100k most frequent words after filtering out all words with less than 20 occurrences or with occurrence in more than 10% of documents.
Using this dictionary we extract bag-of-words feature vectors from the available captions in the MSCOCO and Flickr30k datasets and we compute the *tf-idf* coefficients from the joint corpus. Using the transformed vectors we train a 400 topic LSI model. Textual similarity between two strings is computed as cosine distance between their representations in latent topic space. When comparing text queries, the starting *find shots of* is stripped in preprocessing.

3.4 Classification and regression

For the classification step (used in Runs 1 and 2) we used SVMs with nonlinear RBF and chi-square kernels. The training set consists of "fc7" features described in section 3.2 extracted from the 1957 manually collected images. Optimizing the SVM hyperparameter γ as well as the regularization parameter C is accomplished by gridsearch with values between $1e-4$ and 10 in logarithmic increments. Cross-validation is performed with a stratified 3-fold strategy, and multiclass is enforced through one-vs-rest. The mean value for the the classification score (accuracy) on cross-validation is at around 0.8. Estimator scores are converted to probabilities by using Platt's rule. The estimated probabilities from the 2 classifiers (Gaussian and chi-square) are combined into one probability score for each shot using the "or rule"

$$p_{shot} = 1 - (1 - p_{RBF}) * (1 - p_{\chi^2}) \qquad (1)$$

For the regression step we used random forest and support vector regression. The training data consisted of "fc7" features obtained from all the images in the MSCOCO and Flickr30k datasets. The regression target values represent text similarity (as defined in section 3.3) between the image's human annotation and the 'payload' part of each query (i.e., with the 'find shots of' stripped). For most captioned images there were several annotations per image: we max-pool the similarity value at image level. Because of time and memory concerns we were only able to train with a linear kernel implementation based on LIBLINEAR.

Table 1: Mean extended inferred average precision per run

Run	Submission Type	Name	mAP
1	Manually-assisted	Fuse all	0.044
2	Manually-assisted	Exemplaries	0.043
3	Fully automatic	Autotraining	0.004
4	Fully automatic	InvertedTalk	0.004

3.5 Fusion

Submission scores for runs 2,3 and 4 have been computed based on late (score) fusion. Given the lack of any validation data and the assumed low occurrence rate of true examples we decided to simply assign a weight of 1 to all features participating in weighted fusion.

4. RESULTS

Table 2 shows the detailed results of four submitted runs.

In the manually-assisted runs, the vitrivr system ranked overall at position 9 (run 1) and position 11 (run 2) out of 22, respectively. In the fully automatic runs, on the other hand, the vitrivr system ranked at position 25 (run 3) and position 26 (run 4) out of 30.

5. CONCLUSION

It it not surprising that the manually-assisted runs largely outperformed the automatic runs, since the former relied on manually collected examples and classification, while the latter depended on the quality of caption retrieval for providing regression targets.

Results of the manual runs have confirmed that established techniques used in previous TRECVID SIN editions — higher order features from CNN upper layers combined with discriminative classifiers like SVMs — are still relevant in the AVS context.

One of the possible causes for the modest performance of the automated runs is in the unreliability of the LSI model's ranking of the captions. For example, in topic 524 all captions containing 'beard' were higher ranked than all captions matching 'white robe'. Even with perfect transfer, this imbalance would manifest directly in the final ranking of the shots. Given the submission list is limited at 1k, the performance consequently degrades.

A serious challenge compared with previous TRECVID editions was the lack of training and validation data for the imposed topics. One direct consequence was the inability to fine-tune parameters for the classification and fusion components. However, even with simple score summing used as fusion, the MAP improvements from Run4 up to Run1 are confirmed as significant by the TRECVID randomized test.

6. ACKNOWLEDGMENTS

This work was partly supported by the Swiss National Science Foundation (SNSF, contract no. 20CH21_151571) in the context of the IMOTION project.

[2] https://github.com/BVLC/caffe/tree/master/models/bvlc_reference_caffenet

[3] https://github.com/pluskid/Mocha.jl

[4] https://github.com/karpathy/neuraltalk2

[5] available at https://dumps.wikimedia.org/enwiki/20160701/

Table 2: Mean extended inferred average precision per query and run

Task	Run 1	Run 2	Run 3	Run 4
501	**0.02**	0.01	0.00	0.00
502	0.16	**0.18**	0.00	0.00
503	**0.20**	0.17	0.01	0.00
504	0.06	0.06	0.01	0.00
505	0.01	0.01	0.00	0.00
506	0.23	**0.28**	0.00	0.00
507	0.16	**0.18**	0.00	0.00
508	**0.01**	0.00	0.00	0.00
509	**0.09**	0.04	0.03	0.00
510	0.00	0.00	0.00	0.00
511	0.01	0.01	0.00	0.00
512	0.01	**0.03**	0.00	0.00
514	**0.01**	**0.01**	0.00	0.00
514	**0.01**	**0.01**	0.00	0.00
515	**0.01**	0.00	0.00	0.00
516	0.00	0.00	0.00	0.00
517	**0.01**	**0.01**	0.00	0.00
518	**0.05**	0.03	0.02	0.02
519	0.03	0.01	0.03	**0.07**
520	**0.04**	**0.04**	0.00	0.00
521	0.00	**0.01**	0.00	0.00
522	**0.05**	0.04	0.01	0.00
523	**0.02**	**0.02**	0.00	0.00
524	**0.01**	**0.01**	0.00	0.00
525	**0.01**	0.00	0.00	0.00
526	**0.01**	**0.01**	**0.01**	0.00
527	0.00	0.00	0.00	0.00
528	0.11	**0.12**	0.00	0.01
529	0.00	**0.02**	0.00	0.00
530	0.00	0.00	0.00	0.00
Median	**0.012**	0.011	0.01	0.00
Average	**0.044**	0.043	0.004	0.004

7. REFERENCES

[1] G. Awad, J. Fiscus, M. Michel, D. Joy, W. Kraaij, A. F. Smeaton, G. Quéenot, M. Eskevich, R. Aly, and R. Ordelman. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID 2016*. NIST, USA, 2016.

[2] I. Giangreco and H. Schuldt. ADAMpro: Database support for big multimedia retrieval. *Datenbank-Spektrum*, 16(1):17–26, 2016.

[3] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[4] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] 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 *European Conference on Computer Vision*, pages 740–755. Springer, 2014.

[6] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649, 2015.

[7] L. Rossetto, I. Giangreco, and H. Schuldt. Cineast: a multi-feature sketch-based video retrieval engine. In *Multimedia (ISM), 2014 IEEE International Symposium on*, pages 18–23. IEEE, 2014.

[8] L. Rossetto, I. Giangreco, C. Tanase, and H. Schuldt. vitrivr: A flexible retrieval stack supporting multiple query modes for searching in multimedia collections. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 1183–1186. ACM, 2016.

[9] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.