# Enhanced Retrieval and Browsing in the IMOTION System

Luca Rossetto[1], Ivan Giangreco[1], Claudiu Tănase[1],
Heiko Schuldt[1], Stéphane Dupont[2], and Omar Seddati[2]

[1] Databases and Information Systems Research Group,
Department of Mathematics and Computer Science, University of Basel, Switzerland
{luca.rossetto|ivan.giangreco|c.tanase|heiko.schuldt}@unibas.ch
[2] Research Center in Information Technologies, Université de Mons, Belgium
{stephane.dupont|omar.seddati}@umons.ac.be

**Abstract.** This paper presents the IMOTION system in its third version. While still focusing on sketch-based retrieval, we improved upon the semantic retrieval capabilities introduced in the previous version by adding more detectors and improving the interface for semantic query specification. In addition to previous year's system, we increase the role of features obtained from Deep Neural Networks in three areas: semantic class labels for more entry-level concepts, hidden layer activation vectors for query-by-example and 2D semantic similarity results display. The new graph-based result navigation interface further enriches the system's browsing capabilities. The updated database storage system ADAM$_{pro}$ designed from the ground up for large scale multimedia applications ensures the scalability to steadily growing collections.

## 1 Introduction

In this paper we introduce the 2017 version of the IMOTION system which is the third iteration (after [13] and [11]) of the system participating in the Video Browser Showdown [2].

We provide a brief overview of the overall architecture of the system in Section 2, and elaborate in greater detail on the improvements made since the previous version in Section 3. Section 4 concludes.

## 2 The IMOTION System

### 2.1 Overview

The IMOTION system is a sketch-based video retrieval system which supports a large variety of query paradigms, including query-by-sketch, query-by-example, query-by-motion and querying using semantic concepts. It allows to search using multiple query containers, e.g., a still image, a user-provided sketch, the specification of motion via flow fields or by denoting a semantic concept. The IMOTION system is built in a flexible and modular way and can easily be extended to support further query modes or feature extractors.

### 2.2 Architecture

The 2017 IMOTION system is based on the ADAM$_{pro}$ database [3] and the Cineast retrieval engine [12] which are both part of the vitrivr[3] open-source content-based multimedia retrieval stack [14]. The IMOTION system has a custom browser-based front end which communicates with the storage and retrieval back-end via a web server which also serves the static content such as videos and preview images. Figure 1 shows an overview of the architecture of the IMOTION system.

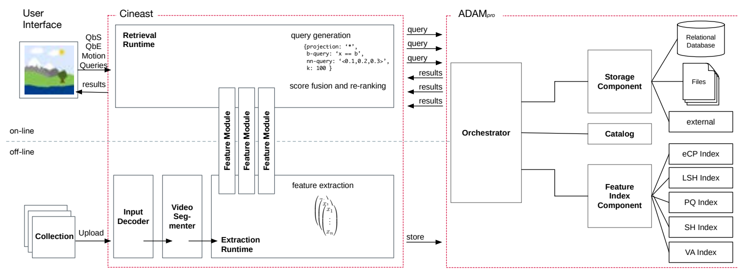


**Fig. 1.** Architectural overview of the IMOTION system.

# 3 New Functionality

### 3.1 Concept Detection

Since the last edition, we have expanded the set of semantic features supported by IMOTION. All these features are based on Deep Neural Network classifiers:

- We have extracted semantic categories representing entry-level labels of environments from the Places2 dataset. Classification was performed using the pre-trained VGG16-places365 network [18].
- We have trained image-level classifiers for the 80 classes of the MS COCO Detection challenge [9]. The feature data is obtained from the last fully connected layer ("fc7") of a VGG convolutional network. The model is trained on the MS COCO train2014 data and it learns the 80 labels independently using multinominal logistic regression.
- We kept the 325 semantic entry-level categories obtained from $n$-grams from last year [11].

Given the participation in this year's TRECVID Ad Hoc Search task[4], which also operated on the IACC.3 data, we integrated the result scores for our estimated best run into the search engine. We have extended the list of 30 AVS

[3] https://www.vitrivr.org/
[4] http://www-nlpir.nist.gov/projects/tv2016/tv2016.html#avs

textual queries with several queries we consider useful for browsing e.g., "shots with two people", "shots showing cartoons", etc.

As in our previous system, we use multiple ConvNets for feature extraction and object/action recognition. We replaced the temporal ConvNet trained on dense optical flow maps with ConvNets that are able to recognize visual actions that may be detected from single images. In order to train these ConvNets, we used the two databases Stanford 40 [17] with 40 categories of actions and COCO-a with 140 categories [10].

We also use a modified version of the DenseCap [7] language model (LM). We use a beam search approach in order to keep multiple results at each generated word. We hence end up with a number of alternatives sentences for each region of interest. From these sentences, we recover a set of words corresponding to objects and attributes. We also use downsampled (bilinear sampling) features extracted with DenseCap ConvNet. This ConvNet was trained on the Visual Genome [8] dataset.

### 3.2 Semantic Class Selection

As with the previous version of the system, one supported query mode is to search for instances of detected semantic concepts. In the 2016 IMOTION system [11] we implemented the interface for the selection of these concepts as a list of icons which could be added to a canvas via drag and drop. Figure 2 shows an example of this UI element. The new selection interface for VBS 2017, depicted in Figure 3, uses a text box with an auto-complete feature to select semantic classes. Every class adds a weight slider by which the importance of this class with respect to the query can be specified.

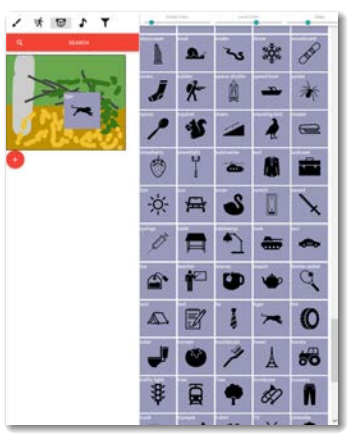

**Fig. 2.** Semantic class selection in the 2016 IMOTION system.

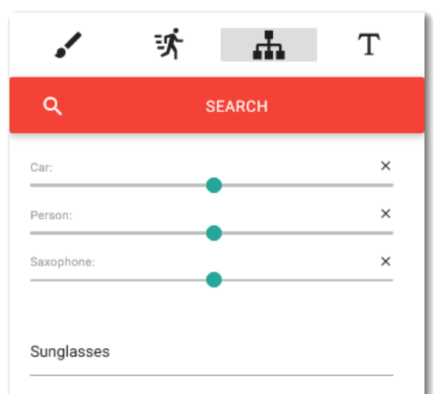


**Fig. 3.** New semantic class selection in the 2017 IMOTION system.

### 3.3 Result Presentation and Browsing

In addition to the existing querying capabilities, for the 2017 version of the system we put additional emphasis on exploratory search and browsing capabilities. In a manner similar to several of the 2016 VBS systems (e.g., [1]), we have implemented a similarity-based navigation interface. The new interface allows to navigate through the resulting grid by panning and zooming as it places visually and semantically similar results close to each other.

### 3.4 Text-based retrieval

At VBS 2017, we use traditional text retrieval based on Lucene to search in the text extracted from the ASR (as provided with the video data), and captions extracted from the keyframes using DenseCap [7].

### 3.5 ADAM$_{\mathbf{pro}}$

In the most current version, the IMOTION system uses the new ADAM$_{pro}$ database. The ADAM$_{pro}$ database [3] is geared towards offering storage and retrieval capabilities for multimedia objects and the corresponding metadata. To this end, it supports both Boolean retrieval and $k$ nearest neighbour similarity searches in the vector space retrieval model and is particularly tailored to support large multimedia collections. ADAM$_{pro}$ comes with various index structures that are very different in their nature: Locality-Sensitive Hashing [5] and Spectral Hashing [16] are hash-based methods and form together with Product Quantization [6] and extended Cluster Pruning (eCP) [4] a group of indexes which support a rather coarse retrieval which can be executed very quickly, however suffers from the fact that it may miss result candidates as they are pruned by mistake from the candidate list. The Vector Approximation-File (VA-File) index [15], on the other hand, may degenerate to a sequential scan in worst case; however, it will not prune by mistake a true result candidate. Finally, ADAM$_{pro}$ supports sharding a collection to multiple nodes to increase the retrieval efficiency.

## 4 Conclusions

The 2017 version of the IMOTION system has received significant upgrades over previous versions in both indexing and browsing. Compared to last year's version, we have tripled the number of semantic classes and improved the class selection mechanism. In agreement with video browsing state of the art, the results browsing interface features semantic-based arrangement, which is supposed to significantly reduce the interaction overhead for browsing and near-hit search. Finally, the new version of IMOTION is backed up by the new ADAM$_{pro}$ storage system, which comes with a large variety of indexing structures to decrease query latency.

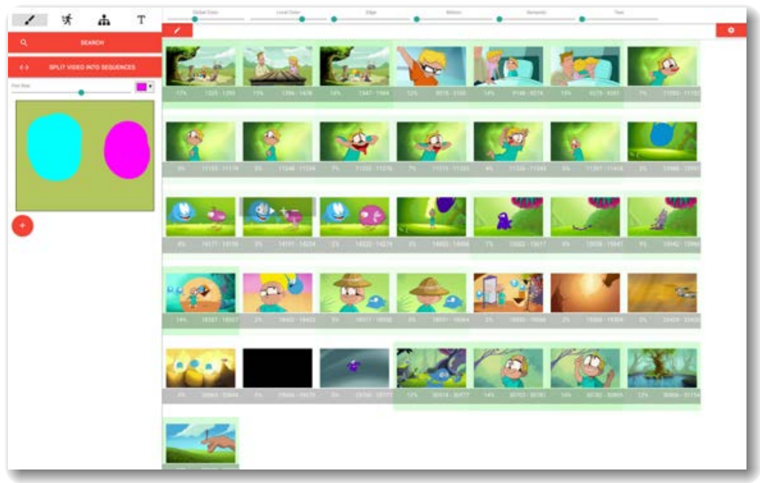


**Fig. 4.** Screenshot of the 2017 IMOTION system UI.

## Acknowledgements

This work was partly supported by the Chist-Era project IMOTION with contributions from the Belgian Fonds de la Recherche Scientifique (FNRS, contract no. R.50.02.14.F) and the Swiss National Science Foundation (SNSF, contract no. 20CH21_151571).

## References

1. Kai Uwe Barthel, Nico Hezel, and Radek Mackowiak. Graph-based browsing for large video collections. In *International Conference on Multimedia Modeling*, pages 237–242. Springer, 2015.
2. Claudiu Cobârzan, Klaus Schoeffmann, Werner Bailer, Wolfgang Hürst, Adam Blažek, Jakub Lokoč, Stefanos Vrochidis, Kai Uwe Barthel, and Luca Rossetto. Interactive video search tools: a detailed analysis of the video browser showdown 2015. *Multimedia Tools and Applications*, pages 1–33.
3. Ivan Giangreco and Heiko Schuldt. ADAMpro: Database support for big multimedia retrieval. *Datenbank-Spektrum*, 16(1):17–26, 2016.
4. Gylfi Gudmundsson, Björn Jónsson, and Laurent Amsaleg. A large-scale performance study of cluster-based high-dimensional indexing. In *Proc. Int. Ws. on very-large-scale multimedia corpus, mining and retrieval (VLS-MCMR 2010)*, pages 31–36, Firenze, Italy, 2010. ACM.
5. Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proc. Symp. on the Theory of Computing*, pages 604–613, Dallas, Texas, USA, 1998. ACM.
6. Herve Jegou, Matthijs Douze, and Cordelia Schmid. Product quantization for nearest neighbor search. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 33(1):117–128, 2011.

7. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
8. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv preprint arXiv:1602.07332*, 2016.
9. T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár. Microsoft COCO: Common Objects in Context. *ArXiv e-prints*, May 2014.
10. Matteo Ruggero Ronchi and Pietro Perona. Describing common human visual actions in images. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC 2015)*, pages 52.1–52.12. BMVA Press, September 2015.
11. Luca Rossetto, Ivan Giangreco, Silvan Heller, Claudiu Tănase, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, Ozan Can Altıok, and Yusuf Sahillioğlu. IMOTION – searching for video sequences using multi-shot sketch queries. In *International Conference on Multimedia Modeling*, pages 377–382. Springer, 2016.
12. Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. Cineast: a multi-feature sketch-based video retrieval engine. In *Multimedia (ISM), 2014 IEEE International Symposium on*, pages 18–23. IEEE, 2014.
13. Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, and Yusuf Sahillioğlu. IMOTION – a content-based video retrieval engine. In *MultiMedia Modeling*, pages 255–260. Springer, 2015.
14. Luca Rossetto, Ivan Giangreco, Claudiu Tanase, and Heiko 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.
15. Roger Weber, Hans-Jörg Schek, and Stephen Blott. A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proc. Int. Conf. on Very Large Data Bases (VLDB 1998)*, pages 194–205, New York, USA, 1998.
16. Yair Weiss, Antonio Torralba, and Robert Fergus. Spectral hashing. In *Proc. Annual Conference on Neural Information Processing Systems (NIPS 2008)*, pages 1753–1760, Vancouver, Canada, 2008.
17. Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *2011 International Conference on Computer Vision*, pages 1331–1338. IEEE, 2011.
18. Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014.