Multimodal Video Retrieval with the 2017 IMOTION System

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

ABSTRACT

The IMOTION system is a multimodal content-based video search and browsing application offering a rich set of query modes on the basis of a broad range of different features. It is able to scale with the size of the collection due to its underlying flexible polystore called ADAM$_{pro}$ and its very effective retrieval engine Cineast, optimized for multi-feature fusion. IMOTION is simultaneously geared towards precision-focused searches, i.e., known-item search with image or text queries, and recall-focused, exploratory searches. In this demo, we will present the 2017 IMOTION system deployed on the IACC.3 collection consisting of 600 hours of Internet Archive video, which was also used in the TRECVID 2016 Ad-Hoc Video Search and in the 2017 Video Browser Showdown (VBS) challenge in which IMOTION ranked first. Conference attendees will have the chance to interact with the 2017 IMOTION system and quickly solve various retrieval tasks.

KEYWORDS

Content-based retrieval; Video retrieval; Multi-modal retrieval; Indexing; Video browsing

ACM Reference format:
Luca Rossetto Ivan Giangreco Claudiu Tănase Heiko Schuldt. 2017. Multimodal Video Retrieval with the 2017 IMOTION System. In *Proceedings of ICMR '17, Bucharest, Romania, June 6–9, 2017,* 4 pages.
DOI: http://dx.doi.org/10.1145/3078971.3079012

1 INTRODUCTION

Information retrieval technologies are key enablers of a range of applications in an environment where finding the right information and data becomes critical in many sectors, for efficient decision-making, research, or creative thinking. Multimedia content deserves a particular treatment given its unstructured (non-symbolic) and hidden (semantic gap) meaning to the computer. This calls for research on the way it can be indexed, queried, and stored efficiently.

The IMOTION system [14] is a content-based video search engine that provides fast and intuitive *known item search* in large video collections. In this paper, we pay special attention to the extensions and improvements we brought to the system after the publication of [16]. In this demo, we showcase the IMOTION system with all the available features and query modalities using a collection of 600 hours of video (IACC.3). We present the 2017 IMOTION system [16] which won the 2017 iteration of the Video Browser Showdown (VBS) [4]. Conference attendees will be able to interactively use the system and explore the collection using all available modes.

This paper is structured as follows: Section 2 introduces the system architecture; Section 3 presents the capabilities of the system from the features, the query modalities and the user interaction point of view. The demo plan is discussed in Section 4. Section 5 presents related work and Section 6 concludes.

Table 1: Overview of feature data sources and query modes

Feature		Query Modality
Low-level	Image	Query-by-Sketch, Query-by-Example, Relevance Feedback
	Motion	Query-by-Motion
Mid-level	On-screen Text	Text-based Retrieval
	Spoken Audio	
High-level	Automated Image Captioning	
	User-provided Information	
	Semantic Concepts	Weighted Boolean retrieval

2 SYSTEM ARCHITECTURE

The IMOTION system is an extension of the vitrivr open source retrieval stack [15] and hence follows its architecture. It shares the polystore (ADAM$_{pro}$) [6] and the retrieval engine (Cineast) [13] with vitrivr while extending both, mainly by making use of text-based retrieval capabilities which are based on Apache Solr[1], and by using a custom user interface (see Figure 1) optimized for the challenges of the VBS 2017 competition.

3 IMOTION SYSTEM CAPABILITIES

In this section, we present in detail the system capabilities and the broad range of retrieval modes of the system. We give an overview of the features and query modalities summarized in Table 1.

3.1 Features

Low-level Image Features The color category contains low-level image features such as global and localized color histograms and regional color aggregations such as the color

ICMR '17, Bucharest, Romania

978-1-4503-4701-3/17/06…$15.00.
DOI: http://dx.doi.org/10.1145/3078971.3079012

[1] https://lucene.apache.org/solr/

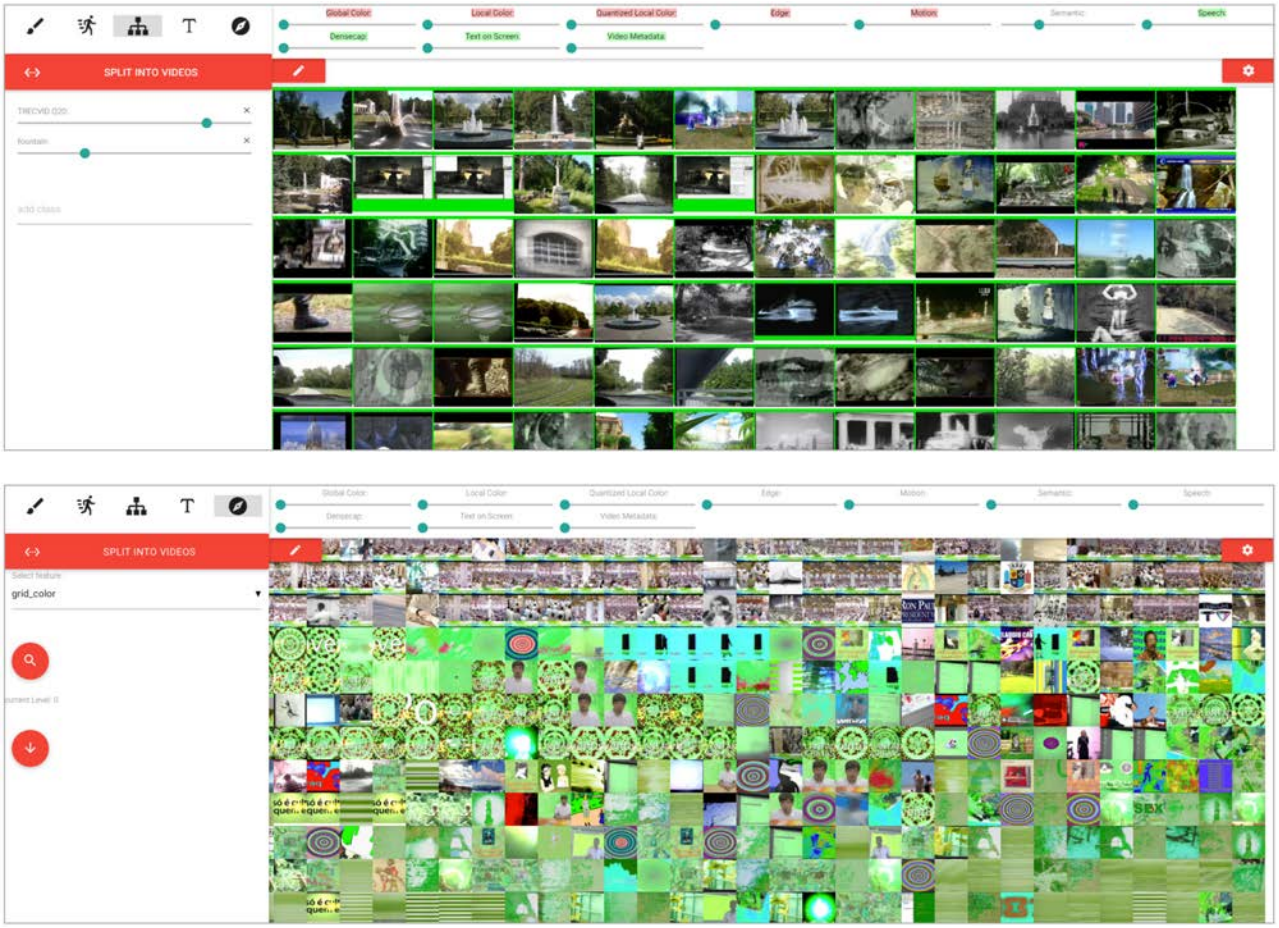

Figure 1: Screenshots of the GUI of the 2017 IMOTION system. Top: known-item search GUI. Bottom: exploratory search GUI.

layout descriptor. Similarly, the edge feature category contains features concerned with localized directional and non-directional edge histograms.

Low-level Motion Features To describe motion, localized directional histograms of sparse trajectories are used. During extraction, background–foreground separation is done by stabilizing the frame with a homographical model and the foreground region is estimated by thresholding the result of a spatio-temporal Gaussian blurring convolution.

OCR Since many videos contain sequences prominently featuring text on screen, we apply the Tesseract [18] optical character recognition system to every keyframe. For videos segmented into five or fewer shots, we sample the videos additionally at a 10 second interval. Because the output of the OCR is non-predictable for images which do not contain any text, it is filtered using the ratio of special characters to regular characters in order to reduce the amount of noise produced.

Speech-to-Text The metadata provided with the video collection contains among other things the output of an automatic speaker recognition (ASR) process [5, 9] which tries to detect all speakers within a video and timestamp every spoken word. We use this ASR data and transform it into artificial close captions by thresholding the detected words by confidence value and by mapping them to the boundaries of the shots. These captions are then aggregated into one document per shot such that each document contains all words spoken during the shot as well as the ones from the previous and the subsequent shot. This windowing is done in order to increase the chances of retrieving sequences in over-segmented video regions.

Automated Image Captioning We apply the DenseCap [7] image captioning neural network to all keyframes of the collection to produce one or several simple textual descriptions of the depicted scene. These captions are then grouped into one document per keyframe and subsequently loaded into Solr. The captioning is able to label a wider range of objects since the network is not trained to recognize objects directly but rather to pair textual descriptions to images and is hence not limited to a fixed set of classes.

User-provided Information For each video, the collection also contains a file with metadata. Even though these files

are encoded in XML, the structure is of little benefit which is why we flatten these documents by removing all XML tags. The remaining text documents are also added to Solr. Since this is the only instance where we have one document per video rather than one per shot, we return one in ten shots of all videos which match a query on this modality.

Semantic Concept Instances The system can filter results by various semantic categories, which are obtained by binarizing prediction outputs of a total of 655 external classifiers. We apply various object and concept detectors on keyframes of the video collection and store the top 600 instances per detector. A custom network [12] is used to detect 325 common objects, defined by entry-level nouns (e.g., 'dog'). In addition, we added 365 labels corresponding to classes in the Places2 [22] dataset representing environments (e.g., 'amusement park') and 80 object category labels learned from human annotation in the MSCOCO [10] detection dataset (e.g., 'snowboard'). Scores for the 30 topics we extracted for TRECVID AVS 2016 [20] are also present as categories. Since all these lists are precomputed, we store them directly in the browser-based user interface to reduce runtime system load on the back-end.

3.2 Query Modalities

The following query modalities of IMOTION can be used either sequentially (for incremental query refinement) or concurrently (combined search).

Query-by-Sketch To be able to specify queries visually, the system offers a sketch canvas. The user has the possibility to input a rough visual representation of the desired content which is then processed by the same visual features which were used to process the videos.

Query-by-Example In contrast to the *Query-by-Sketch* modality, Query-by-Example does not use a user created object for querying but rather the internal description of an already known object. This way a user has the possibility to extend upon previously retrieved results.

Relevance Feedback In extension to *Query-by-Example*, relevance feedback enables a user to construct a query based on multiple previously retrieved results by labeling them either as relevant or irrelevant. The system then also uses the internal representations of these positive and negative examples to construct a new result.

Query-by-Motion Similar to *Query-by-Sketch*, the user has the possibility to specify flow-fields which roughly describe the motion within the video sequence. The interface allows switching between foreground and background motion sketching. Foreground flow fields model prominent moving objects and background flow fields model camera movement or movements of small objects. It should be noted that an empty motion canvas essentially means querying for still sequences.

Text-based Retrieval The text retrieval modality allows for querying using keywords. It makes use of the text-based retrieval capabilities provided by Apache Solr, i.e., computing the relevance using term frequency/inverse document frequency (TF-IDF).

Weighted Boolean Retrieval For the *Semantic Concept Instances*, we use pre-computed lists of relevant shots per topic. All list elements have a score which was determined by the rank of the detector output, the score ramps in the interval [0.0, 1.0] over the length of the list. By selecting a concept as relevant and thereby adding all elements of its corresponding list to the result set, the UI also adds a weight slider for the concept as shown in Figure 1 on the left. Using these sliders, the user can specify the relevance of a topic with respect to the others which provides more overall flexibility than basic Boolean retrieval.

3.3 User Interaction

3.3.1 Precision-focused searches (Known-item Search). For precision-focused searches, i.e., known-item searches, the IMOTION system provides a search interface as shown in Figure 1: query results are displayed with a thumbnail of the shot and a border which indicates its relevance to the query. The sliders on top control the weights of the various similarity measures. Changing the value of one of these sliders updates the scores of the individual results and re-orders them correspondingly in real time. In cases where the correct result is difficult to identify based on this grid representation, the user has the option to re-group the results by video. The results are then grouped chronologically within a video (horizontally) while the videos are ordered by maximum shot score (vertically). For every shot, the user also has the option to load the previews of the surrounding shots or of all shots from the same video in order to be able to gain a better overview of the entire video. It is also possible to start the video playback from every shot at the relevant position.

3.3.2 Recall-focused searches (Exploratory search). In order to complement the Query-by-Sketch/Example approach with exploratory navigation, the system offers multiple 2D-embeddings of the collection, based on different similarity metrics. These embeddings manifest in the UI as a separate view whereupon keyframes are placed in a large complete 572×572 grid. The embedding method assigns the keyframes to grid positions such that distance relations in a given feature space are best preserved.

An important feature is the seamless two-way transfer between search and embedding mode. Any thumbnail from the results pane can be instantly centered in the 2D map in order to allow exploration of its neighbors. Conversely, every object of the 2D map can be used as query, with optional refinement.

The 2D embeddings are obtained by in three steps: a sampling is first reprojected to 2D space using t-distributed Stochastic Neighbor Embedding [11], then mapped to coarse grid coordinates using optimal linear assignment (we use an implementation of the Jonker-Volgenant algorithm [8]), 2D positions of the full set are then estimated using Gaussian RBF interpolation and finally mapped to fine grid coordinates by greedy assignment.

4 DEMONSTRATION

For this demonstration conference attendees will have the chance to perform known-item searches with one of the pre-selected visual or textual queries, try AVS-style topics or simply explore the video collection with the 2017 IMOTION system. The presented system will be in the setup used for the 2017 Video Browser Showdown, using the IACC.3 video collection of over 600 hours from the Internet Archive.

5 RELATED WORK

Content-based interactive video retrieval has attracted a relatively small but dedicated research community. The Video Browser Showdown [4] is an annual competitive live benchamrking event where particpating teams are asked to submit in a limited time (no more than 5 minutes) results matching an ad-hoc video query. These queries have been only of the Known Item Search variety (one unique correct result in the entire collection) until the 2017 edition of VBS, where more exploratory AVS queries have been introduced. The AVS queries (e.g., "Find all shots of women wearing glasses"), as well as the pre-defined video collection are based on the TRECVID [1] 2016 Ad-Hoc Video Search task. This demo uses some results from the IMOTION team's participation in this task [20].

The field of content-based video retrieval without the interactivity constraint is more prominent in the literature. Two essential benchmarks for such search systems are the TRECVID and MediaEval challenges. Without the response time restrictions, content analysis systems in this field are able to make use of computationally heavier methods such as Deep Neural Networks for various classification tasks [21]. An important recent development in real-time search borrowed from this body of work consists of tagging the collection with semantic classifiers or using hidden layer activations as features in the pre-processing phase [20].

Recent high-performing approaches in video browsing revolve around retrieval of simplified sketches (e.g., by using simple color signatures [3]) and displaying the collection in a more informative way (e.g., using a graph-based keyframe arrangement for browsing [2]). A more in-depth sketch analysis where deep semantic classifiers are employed for sketch auto-completion has been demonstrated in earlier work [19]. For a comprehensive review of video search or browsing methods and systems, readers can refer to [17].

6 CONCLUSION

In this paper, we have presented the 2017 edition of the IMOTION system and we have shown how it can be used in an interactive demonstration for retrieving video sequences using a large variety of features and query modalities.

ACKNOWLEDGEMENTS

This work was partly supported by the Chist-Era project IMOTION with contributions from the Swiss National Science Foundation (SNSF, contract no. 20CH21_151571).

REFERENCES

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

[2] Kai Uwe Barthel, Nico Hezel, and Radek Mackowiak. 2016. Navigating a graph of scenes for exploring large video collections. In *Proceedings of the 22nd International Conference on Multimedia Modeling (MMM 2016)*. Springer, Miami, FL, USA, 418–423.

[3] Adam Blažek, Jakub Lokoč, Filip Matzner, and Tomáš Skopal. 2015. Enhanced signature-based video browser. In *Proceedings of the 21st International Conference on Multimedia Modeling (MMM 2015)*. Springer, Sydney, NSW, Australia, 243–248.

[4] Claudiu Cobârzan, Klaus Schoeffmann, Werner Bailer, Wolfgang Hürst, Adam Blažek, Jakub Lokoč, Stefanos Vrochidis, Kai Uwe Barthel, and Luca Rossetto. 2016. Interactive video search tools: A detailed analysis of the video browser showdown 2015. *Multimedia Tools and Applications* (2016), 1–33.

[5] Jean-Luc Gauvain. 2010. The Quaero Program: Multilingual and Multimedia Technologies. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2010)*.

[6] Ivan Giangreco and Heiko Schuldt. 2016. ADAM$_{pro}$: Database Support for Big Multimedia Retrieval. *Datenbank-Spektrum* 16, 1 (2016), 17–26.

[7] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully Convolutional Localization Networks for Dense Captioning. In *Proc. of the International Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. IEEE.

[8] R. Jonker and A. Volgenant. 1987. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing* 38, 4 (1987), 325–340.

[9] Lori Lamel. 2012. Multilingual Speech Processing Activities in Quaero: Application to Multimedia Search in Unstructured Data. In *Proc. of the International Conference on Human Language Technologies -– The Baltic Perspective*. 1–8.

[10] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Proc. of the European Conference on Computer Vision (ECCV 2014)*. Springer, 740–755.

[11] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[12] Luca Rossetto, Ivan Giangreco, Silvan Heller, Claudiu Tănase, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, Ozan Can Altıok, and Yusuf Sahillioğlu. 2016. IMOTION–Searching for Video Sequences Using Multi-Shot Sketch Queries. In *Proceedings of the 22nd International Conference on Multimedia Modeling (MMM 2016)*. Springer, Miami, FL, USA, 377–382.

[13] Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2014. Cineast: a multi-feature sketch-based video retrieval engine. In *Proceedings of the International Symposium on Multimedia (ISM 2014)*. IEEE, Taichung, Taiwan, 18–23.

[14] Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, Omar Seddati, T. Metin Sezgin, and Yusuf Sahillioglu. 2015. IMOTION - A Content-Based Video Retrieval Engine. In *Proceedings of the 21st International Conference on MultiMedia Modeling (MMM 2015) (Lecture Notes in Computer Science, Vol. 8936)*. Springer, Sydney, NSW, Australia, 255–260.

[15] Luca Rossetto, Ivan Giangreco, Claudiu Tanase, and Heiko Schuldt. 2016. vitrivr: A Flexible Retrieval Stack Supporting Multiple Query Modes for Searching in Multimedia Collections. In *Proceeding of the International Conference on Multimedia (ACM MM 2016))*. ACM, Amsterdam, The Netherlands, 1183–1186.

[16] Luca Rossetto, Ivan Giangreco, Claudiu Tănase, Heiko Schuldt, Stéphane Dupont, and Omar Seddati. 2017. Enhanced Retrieval and Browsing in the IMOTION System. In *Proceedings of the 23rd International Conference on Multimedia Modeling (MMM 2017)*. Springer, Reykjavik, Iceland, 469–474.

[17] Klaus Schoeffmann, Frank Hopfgartner, Oge Marques, Laszlo Boeszoermenyi, and Joemon M. Jose. 2010. Video Browsing Interfaces and Applications: a Review. *SPIE Reviews (Online)* 1 (March 2010), 1–35.

[18] Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Proceedings of the International Conference on Document Analysis and Recognition (ICDAR 2007)*. IEEE, 629–633.

[19] Claudiu Tanase, Ivan Giangreco, Luca Rossetto, Heiko Schuldt, Omar Seddati, Stephane Dupont, Ozan Can Altiok, and Metin Sezgin. 2016. Semantic Sketch-Based Video Retrieval with Autocompletion. In *Proc. of the International Conference on Intelligent User Interfaces (IUI 2016)*. ACM, Sonoma, CA, USA, 97–101.

[20] Claudiu Tanase, Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, and Omar Seddati. 2016. The IMOTION System at TRECVID 2016: The Ad-Hoc Video Search Task. (2016).

[21] Kazuya Ueki, Kotaro Kikuchi, Susumu Saito, and Tetsunori Kobayashi. 2016. Waseda at TRECVID 2016: Ad-hoc Video Search. (2016).

[22] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. 2016. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055* (2016).