VideoSketch

Innovative Query Modes for Searching and Manipulating Video through Motion and Sound (eNTERFACE'2015 – 10/08-04/09/2014)

Investigators

- Stéphane Dupont, Omar Seddati, Christian Frisson, Gueorgui Pironkov, Saïd Mahmoudi (University of Mons)
- Ivan Giangreco, Luca Rossetto, Claudiu-Ioan Tanase, Heiko Schuldt (University of Basel)
- Yusuf Sahillioglu, Metin Sezgin (KOC University Istanbul)

Abstract

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 sector, for efficient decision-making, research, and 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. Despite the numerous recent research advances, we are far from there yet, as can be seen through the increasing number of challenging benchmarks and competitions related to MIR (Multimedia Information Retrieval). In this context, the study of query modes beyond keywords or text is of particular interest. Here, we will focus on audiovisual databases (possibly augmented with motion information) and develop an integrated prototype enabling to make use of content analysis and recognition through machine learning, multiple query modes (symbolic and non-symbolic enabling to specify the visual as well as audio characteristics of searched video shots), and a scalable database back-end. It will be evaluated in a known-item search task.

Project objectives (max 1 page) - providing the rationale for the proposed project.

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 sector, for efficient decision-making, research, and 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. Despite the numerous recent research advances, we are far from there yet, as can be seen through the increasing number of challenging benchmarks and competitions related to MIR (Multimedia Information Retrieval). In this context, the study of query modes beyond keywords or text is of particular interest. Here, we will focus on audiovisual databases (possibly augmented with motion information) and develop an integrated prototype enabling to make use of content analysis and recognition through machine learning, multiple query modes (symbolic and non-symbolic enabling to specify the visual as well as audio characteristics of searched video shots), and a scalable database back-end. It will be evaluated in a known-item search task.

Background information (max 1 page)

a brief review of the related literature, so as to let potential participants prepare themselves for the workshop.

Existing approaches to video retrieval either focus on audio signals, video frames (images), additional metadata (including subtitles), or a combination of these [Gibbon 2008]. QBIC [Flickner 1995] is one of the first systems that successfully combined image and video retrieval by considering colour, shape, texture, sketches, and even sample images. The MIRACLE project [Gibbon 2006] combines multiple types of video search, such as text searches (transcripts, subtitles, closed captions, and speech recognition), visual information (face clustering and scene change detection), and speaker segmentation. VideoQ [Chang 1998] addresses movement extraction and supports motion queries by automated video object segmentation and tracking, and use real-time video queries. Other approaches to motion-based video indexing and retrieval are reported in [Dagtas 2000, Fablet 2002, Su 2007, Sedmidubský 2013]. The most visited video search engines on the Internet, youtube, Google Video, and Yahoo Video still rely on very basic features, mainly text from closed captions, subtitles, or social metadata.

An active area of research is also semantic video retrieval where video scenes are automatically analysed and tagged with information on their semantics [Ballani 2007, Bertini 2007, Snoek 2006], including visual concepts as well as actions displayed in the videos. So far, many classical image-based features have been used and extended [Scovanner 2007] to video content, with classification relying on bag-of-features and support vector machines, or other classification approaches. Dense or sparse optical flows have also been used [Subramanian 2014]. Recently, deep artificial neural networks and in particular convolutional ones have showed state-of-the-art performance [Simonyan 2014, Karpathy 2014]. In general, the use of convolutional neural networks enables classification without having recourse to classical feature extraction schemes. Concept and action recognition is nevertheless challenging given the large

number of spatial and temporal configurations observed in such content. Finally, beside visual information, the audio channels (through voice but also other acoustic sounds and music) may also provide complementary information [Barchiesi 2015].

Of particular interest is the possibility to rely on query modes beyond keywords and navigation interfaces beyond result lists. With this respect, the use of sketches is particularly relevant. In some cases ideas and concepts that are hard to explain can be naturally and concisely described using sketches. In this respect, sketching is a unique modality. The main objective of the Sketch-based Motion Query work package is to build the infrastructure needed for interpreting sketch-based user queries that specify the motion of objects. Sketches may refer directly to objects or events in the videos, or describe movement. The goal will be to use machine learning algorithms to build sketch recognizers capable of interpreting queries put forward n the form of a drawing. In order to understand what kinds of queries lend themselves to sketch-based specification, we plan on conducting semi-structured interviews with potential end-users of the retrieval system. In reference to the state of the art [Chang 1998, Collomosse 2008, Collomosse 2009, Hu 2012], our work will advance the technology by combining sketches with speech or more generally voice.

When collections of video data and associated metadata are large, this necessitate efficient and effective index structures. Multidimensional indexing methods include hashing [Wang 2014]: examples are the Grid File [Nievergelt 1984] or locality-sensitive hashing (LSH) [Gionis 1999]. Other approaches consider a (balanced) tree-based access structure (e.g., R-Tree [Guttman 1984] and R*Trees, DABS-Tree [Böhm 2000] or M-Tree [Ciaccia 1997]). The generalized search tree (GiST) [Hellerstein 1995] subsumes many of the common features of the latter category. The VA-File uses a data structure which is particularly well suited for efficient nearest neighbour search in high-dimensional spaces [Weber 1998], and currently represents a state-of-the-art approach.

Detailed technical description (max 3 pages)

a. Technical description.

The system we will work on is a sketch-based video retrieval engine supporting multiple query paradigms. For vector space retrieval, the system exploits a large variety of low-level image and video features, as well as high-level spatial and temporal features that can all be jointly used in any combination. In addition, it supports dedicated motion features to allow for the specification of motion within a video sequence. For query specification, our system supports query-by-sketch interactions (users provide sketches of video frames), motion queries (users specify motion across frames via partial flow fields), query-by-example (based on images) and any combination of these, and provides support for relevance feedback.

The project will be structured in six main workpackages:

WP1 – Architecture and Specifications

The objective of WP1 is to define the overall architecture of the system to be developed. This includes the different modules needed for feature extraction, query specification (via sketches and speech) and query execution. In order to follow a modular approach, the interfaces between

these components will receive special consideration so that implementations can be replaced without impacting the entire system. This will also allow for a seamless continuation of the development work after the workshop.

WP2 – Data Collection and Annotation

The objective of WP2 is to create a large collection of videos, annotations and relevance judgments for a predefined set of queries for evaluating the system developed in the course of the workshop. For this purpose, a crawler that downloads creative commons / public domain licensed video material together with its corresponding metadata from video pages such as YouTube, vimeo, etc. will be implemented in this work package. In addition, the work package should implement the pre-processing step to bring the videos to a unified format, in particular also in terms of encoding, resolution, etc. Finally, WP2 also includes the extension of the system to log all aspects of the query and the use of the system, i.e., user clicks, queries, search results, system parameters, etc. to ensure an in-depth evaluation of the system.

WP3 – Machine Vision/Audition

The objective of WP3 is to advance the machine learning approaches for feature extraction and classification of video content, as well as mapping between feature extracted from sketch queries and features of the real video content. Both unsupervised and supervised learning will be applied, in particular in order to extract features representative of the semantics of the video shots. Experiments may also be performed to extract such semantic features from the audio channels. High-level semantic features are relevant to support symbolic queries and sketches, while lower level features and mappings will be relevant to non-symbolic ones.

WP4 – Search, Query, and Visualization Interfaces and Modes

The objective of WP4 is to advance the user interfaces enabling to use different modalities for querying the content, to mix and match different modes, and in particular to combine symbolic/semantic ones and non-symbolic ones (closer to the content itself). Query modes specifically addressing the audio channels are also relevant. Visualization of and navigation through retrieved results and the possibility to use those to refine queries (in query-by-example and relevance feedback mechanisms) can also be addressed. It is expected that various interfaces alternatives will be explored by lo-fi prototypes (e.g., paper prototypes).

WP5 - Query execution

To objective of WP5 is to advance the back-end necessary to be able to support the rich multimodal queries. The query execution engine has to be able to process different information modalities simultaneously and combine the outputs of the respective subsets seamlessly. The existing query execution engine will be extended by additional and more capable feature modules, in particular dealing with higher-level semantic representation of visual content. The combination scheme needed for the seamless integration of such intelligent modules as well as the requirements to the underlying storage system will be studied and experimental results will serve as a basis for the subsequent implementation. Additionally, the capability of understanding higher level semantic queries containing descriptions of objects, actions, etc. has to be added to leverage the full potential of the new modules.

WP6 – Evaluation with end-users in a known-item search paradigm

An evaluation campaign will be put in place in order to measure the benefit of different interaction elements in a competitive scenario. Sample queries from the enlarged collection of augmented videos will be shown to a pool of end users for known item search. Each user will be assigned to a variant of the system lacking one component (symbolic/non symbolic sketching, motion, query refinement, etc.). The retrieval performance will be evaluated under different scenarios in accordance to state of the art benchmarks in video search systems such as the Video Browser Showdown [BS 2015]. Results and logs of the user activity will be instrumental in understanding how each query type improves retrieval. Also, sketched queries will enable to complement the data collection prepared in WP2 with additional ground truth information.

b. Resources needed: facility, equipment, software, staff…

Provided:

- Sketch input device (UNIBAS)
- Belgian Chocolate (to thank the end-users and the project participants/developers)

Requested:

- Input/output devices:
    - large sketch input device (wacom cintiq)
    - eye tracking device
    - microphone (vocal input) / webcams (user behaviour study)
    - video projector
    - possibly large scale mocap for investigating 3D queries
- Compute devices:
    - high-performance GPGPU cluster
    - high-performance network switch

c. Project management.

The project will be coordinated by Stéphane Dupont & Omar Seddati who will stay on the workshop site for the whole duration. Other participants will each stay at least two weeks (dates to be determined) and then be available remotely to at least support their contribution (and then possibly upgrade it) to make sure it can be used in the final tests. We intend to be flexible to accommodate for additional inputs from other applicants to the project. Part of it may involve flexible choices in terms of integration approaches (f.i. possibility of data level integration). The development backlog will be defined and reviewed through short meetings every 2 days.

Work plan and implementation schedule (max 1 page)

a tentative timetable detailing the work to be done during the workshop.

The work will be structured and aligned to the four weeks of the eNTERFACE workshop.

Week 1:

- *Specification of the architecture of the system to be developed*. This specification will be revised and extended in the course of the workshop, if necessary.

- *Creation of a video collection*. This includes the implementation of a crawler to harvest videos released under a creative commons / public domain license.
- *Training of NN classifiers*.

Week 2

- *Feature extraction*: extracting semantic features from the video collection created in week 1.
- *Incorporating new feature types:* storing the features in the underlying database of the retrieval system.
- *Adaptation of the retrieval engine / database*, depending on the structure of the semantic features.

Week 3

- *Symbolic sketches for semantic concepts*: performing user studies to find out how to best map the semantic concepts that have been detected to symbolic sketches.
- *Implementing additional query types* that consider symbolic sketches and that map these sketches to semantic concepts.
- Implementing services for s*peech to concept conversion* to support spoken queries.
- *Set-up of server system for running KIS tasks.*

Week 4

- *Running user studies* to evaluate the effectiveness and usability of interfaces for known item video search on the basis of symbolic sketches and spoken queries.
- *Extension and/or refinement of the modules of the system*, if necessary and if indicated by the user studies.

Benefits of the research (max 1 page)

Expected outcomes of the project. Please describe what the tangible results are to be achieved at the end of the Workshop. We insist that all the software components used for the project, and all the software built during the project should be free for use, and available as such to all participants (after the workshop too).

The project will deliver the following items, which will be made available to the community:

1. A collection of videos under creative commons public domain licenses, including annotations and parallel sketch/vocal/natural language queries.
2. The results of a large-scale evaluation using known item search tasks using the collected data, associated sketch queries (symbolic and non-symbolic sketches), as well as vocal and natural language queries. In addition, relevance judgments (this is not available yet!) with ground-truth will be released.
3. Recipes for using open source software on several facets of the project.

The project will also advance the knowledge on:

1. Multi-paradigm query and video known-item search systems back-ends.
2. The design of search interface supporting multiple query modes.

Profile team

a. Leader (with a 1-page-max CV).

Dr. Stéphane Dupont received the PhD degree in EE at the Faculty of Engineering of Mons (Belgium) in 2000. He was post-doctoral associate at ICSI (California) in 2001-2002. In 2002, he joined Multitel (Belgium) to coordinate the speech recognition. In 2008, he joined UMONS (Belgium). He has been involved in numerous transnational projects, as partner and coordinator. His main interests are in speech and audio processing, multimodal HCI, machine learning, and statistical signal processing. He holds 3 international patents and has authored/co-authored over 100 papers in these areas.

b. Staff proposed by the leader (with 1-page-max CVs).

You may propose some members of your future team. If possible, try to avoid having too many people from your group: part of the benefit of eNTERFACE is to let people meet and share experiences from different places, and possibly in different languages.

- Omar Seddati, Christian Frisson, Gueorgui Pironkov, Saïd Mahmoudi (UMONS)
- Ivan Giangreco, Luca Rossetto, Claudiu-Ioan Tanase, Heiko Schuldt (UNIBAS)
- Yusuf Sahillioglu, Metin Sezgin (KOC)

c. Other researchers needed (describing the required expertise for each, max 1 page).

The team is open to other complementary expertise, in particular in the areas of:

- information visualization,
- video summarization,
- human perception and cognition,
- other relevant areas.

References

[Ballani 2007] L. Ballan, M. Bertini, A. Del Bimbo, and W. Nunziati.Soccer players identification based on visual local features.In Proceedings of the 6th ACM international conference on Image and video retrieval, CIVR '07, pages 258-265, New York, NY, USA, 2007. ACM.

[Barchiesi 2015] D. Barchiesi, D. Giannoulis, D. Stowell and M. D. Plumbley. Acoustic Scene Classification. To appear in: IEEE Signal Processing Magazine.

[Bertini 2007] M. Bertini, A. del Bimbo, and C. Torniai. *Soccer video annotation using ontologies extended with visual prototypes*. In: International Workshop on Content-Based Multimedia Indexing (CBMI '07), pp 212-218, June 2007.

[Böhm 2000] C. Böhm and H.-P. Kriegel. *Dynamically optimizing high-dimensional index structures*. In EDBT '2000: Proceedings of the 7th International Conference on Extending Database Technology, pages 36-50, London, UK, 2000. Springer-Verlag.

[Chang 1998] S.-F. Chang; W. Chen, H. J Meng, H. Sundaram, Z. Di. *A fully automated content-based video search engine supporting spatiotemporal queries.* IEEE Transactions on Circuits and Systems for Video Technology, 8(5), pp.602-615, September 1998.

[Ciaccia 1997] P. Ciaccia, M. Patella, and P. Zezula. *M-tree: An effcient access method for similarity search in metric spaces*. In VLDB '97: Proceedings of the 23rd International Conference on Very Large Data Bases, pages 426-435, 1997.

[Collomosse 2008] J.P.Collomosse, G.McNeill, L.Watts. Free-hand sketch grouping for video retrieval. In: Proc. of IEEE International Conference on Pattern Recognition (ICPR 2008), pp.1–4, 2008.

[Collomosse 2009] J. P. Collomosse, G. McNeill, Y. Qian. Storyboard sketches for Content Based Video Retrieval. 12th IEEE International Conference on Computer Vision, pp. 245-252, 2009.

[Dagtas 2000] S. Dagtas, W. Al-Khatib, A. Ghafoor, and R. L. Kashyap. *Models for motion-based video indexing and retrieval*. IEEE Trans. Image Processing, 9(1) pp. 88-101, 2000.

[Fablet 2002] R. Fablet, P. Bouthemy, and P. Perez. *Nonparametric motion characterization using causal probabilistic models for video indexing and retrieval.* IEEE Trans. Image Process., 11(4), pp. 393-407, 2002.

[Flickner 1995] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, and P. Yanker. *Query by Image and Video Content: The QBIC System*. Computer, 28(9):23-32, 1995.

[Gibbon 2006] D. Gibbon, Z. Liu, and B. Shahraray. *The Miracle Video Search Engine*. IEEE CCNC, 2006.

[Gibbon 2008] D. Gibbon and Z. Liu. *Introduction to Video Search Engines*. Springer, 2008.

[Gionis 1999] A. Gionis, P. Indyk, and R. Motwani. *Similarity search in high dimensions via hashing*. In Proc. of VLDB'99, pp. 518-529, September 1999.

[Guttman 1984] A. Guttman. *R-trees: a dynamic index structure for spatial searching*. In SIGMOD '84: Proceedings of the 1984 ACM SIGMOD international conference on Management of data, pp. 47-57, New York, NY, USA, 1984. ACM Press.

[Hellerstein 1995] J. Hellerstein, J. Naughton, and A. Pfefier. *Generalized search trees for database systems*. In VLDB '95: Proceedings of the 21th International Conference on Very Large Data Bases, pages 562-573, 1995.

[Hu 2012] R. Hu, S. James, and J. Collomosse. Annotated free-hand sketches for video retrieval using object semantics and motion. In: Proceedings of the 18th international conference on Advances in Multimedia Modeling (MMM'12), 2012.

[Karpathy 2014] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

[Nievergelt 1984] J. Nievergelt, H. Hinterberger, and K. C. Sevcik. *The grid file: An adaptable, symmetric multikey file structure*. ACM Transactions on Database Systems (TODS), 9(1):38-71, 1984.

[Rossetto 2015] L. Rossetto, I .Giangreco, H. Schuldt, S. Dupont, O. Seddati, M. Sezgin and Y. Sahillioglu. *IMOTION - A Content-based Video Retrieval Engine*. In MMM2015: Proceedings of the 21st MultiMedia Modelling Conference – Video Search Showcase Track, pages 255-260, Sydney, Australia, 2015. Springer-Verlag.

[Sedmidubský 2013] Jan Sedmidubský, Jakub Valcik, Pavel Zezula: A Key-Pose Similarity Algorithm for Motion Data Retrieval. ACIVS 2013: 669-681.

[Scovanner 2007] Scovanner, P., Ali, S., & Shah, M. (2007, September). A 3-dimensional sift descriptor and its application to action recognition. In Proceedings of the 15th international conference on Multimedia (pp. 357-360). ACM.

[Simonyan 2014] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In Advances in Neural Information Processing Systems (pp. 568-576).

[Snoek 2006] C. Snoek, M. Worring, J. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. *The challenge problem for automated detection of 101 semantic concepts in multimedia*. In: Proceedings of the 14th ACM International Conference on Multimedia, pp. 421-430, 2006.

[Su 2007] C.-W. Su, H.-Y.M. Liao, H.-R. Tyan, C.-W. Lin, D.-Y. Chen, and K.-C. Fan. *Motion flow-based video retrieval*. IEEE Transactions on Multimedia, 9(6), pp. 1193-1201,October 2007.

[Subramanian 2014] Subramanian, K., Radhakrishnan, V. B., & Sundaram, S. (2014, April). An optical flow feature and McFIS based approach for 3-dimensional human action recognition. In Intelligent Sensors, Sen-sor Networks and Information Processing (ISSNIP), 2014 IEEE Ninth International Conference on (pp. 1-6). IEEE.

[VBS 2015] Video Browser Showdown http://www.videobrowsershowdown.org/

[Wang 2014] Jingdong Wang, Heng Tao Shen, Jingkuan Song, Jianqiu Ji: Hashing for Similarity Search: A Survey. CoRR abs/1408.2927 (2014)

[Weber 1998] R. Weber, H.-J. Schek, and S. Blott. *A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces*. In Proceedings of the 24rd International Conference on Very Large Data Bases, 1998.