Semantic Sketch-Based Video Retrieval with Autocompletion

Claudiu Tănase
Ivan Giangreco
Luca Rossetto
Heiko Schuldt
University of Basel, CH
c.tanase@unibas.ch
ivan.giangreco@unibas.ch
luca.rossetto@unibas.ch
heiko.schuldt@unibas.ch

Omar Seddati
Stéphane Dupont
Université de Mons, BE
omar.seddati@umons.ac.be
stephane.dupont@umons.ac.be

Ozan Can Altıok
Metin Sezgin
Koç University, Istanbul, TR
oaltiok15@ku.edu.tr
mtsezgin@ku.edu.tr

Abstract
The IMOTION system is a content-based video search engine that provides fast and intuitive *known item search* in large video collections. User interaction consists mainly of sketching, which the system recognizes in real-time and makes suggestions based on both *visual appearance* of the sketch (what does the sketch look like in terms of colors, edge distribution, etc.) and *semantic content* (what object is the user sketching). The latter is enabled by a predictive sketch-based UI that identifies likely candidates for the sketched object via state-of-the-art sketch recognition techniques and offers on-screen completion suggestions. In this demo, we show how the sketch-based video retrieval of the IMOTION system is used in a collection of roughly 30,000 video shots. The system indexes collection data with over 30 visual features describing color, edge, motion, and semantic information. Resulting feature data is stored in ADAM, an efficient database system optimized for fast retrieval.

Author Keywords
Content-based video retrieval; sketch interface

ACM Classification Keywords
H.5.1 [Information Interfaces & Presentation]: Multimedia Information Systems—*Video*; H.5.2 [Information Interfaces & Presentation]: User Interfaces—*Interaction styles*

The IMOTION System

A precursor of the system [9] participated in the 2015 Video Browser Showdown (VBS) [11]. The idea of VBS is that participants are shown a query and they need to retrieve it using their interactive video search system from a large video collection of TV broadcasts. The number of points gained per hit decreases with time, so fast retrieval is encouraged. Also false submissions penalize the maximum number of points gained for that query, which discourages false positives. There are two types of queries: a (visual) short video snippet (around 20 seconds) and a textual query, where only a text is shown describing the contents of the video scene. The video collection for the VBS was of about 100 hours, much larger than the collection in this proposed demo. In spite of these challenges, IMOTION achieved first place on visual querying tasks, second overall.

Introduction

Sketching is a very effective way of representing both abstract and concrete objects. IMOTION is an interactive video retrieval system that offers a sketch-based user interface. It analyzes the sketch from two sides of the semantic gap: at low level, the sketch is considered a low quality visual depiction of the query; at high level, the system tries to guess what it represents and –once the user confirms it– finds all instances in the pre-indexed collection. This exchange translates on the user interface level to sketch completion suggestions. IMOTION has a fast and scalable retrieval backend allowing for query-by-sketch, query-by-example and relevance feedback on a real-world dataset. Browsing the query results is streamlined and efficient, and benefits from live feature weight recombination. A large set of heterogeneous multimodal features describing color, motion, and semantic information is used to characterize data on both query- and collection-side.

Contribution

Our contribution is the IMOTION video retrieval system that jointly supports: i.) a user interface based on a *sketch based autocompletion framework* capable of predicting incomplete sketches of 250 objects; ii.) *Interactive search*, even for very large collection sizes and and category counts; iii.) *Real-time* queries and result update.

System Overview

IMOTION supports three main interaction modes:
i.) for *Query-by-sketch* (QbS) the user sketches the query image with the three available tools: a "pencil" for making object sketches, a "brush" for painting colored sketches and an "arrow" tool for sketching motion trajectories. The pencil tool triggers sketch completion suggestions, while

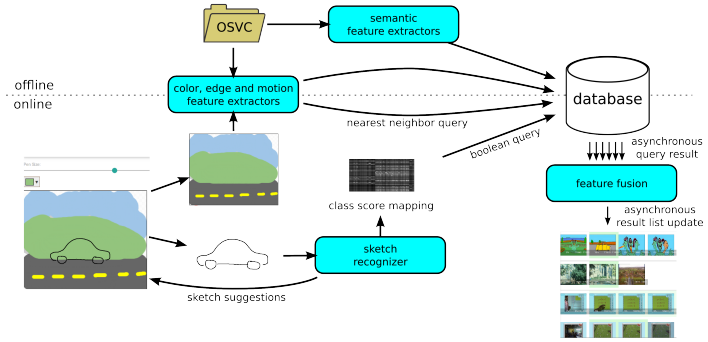

Figure 1: Overview of the IMOTION system architecture.

the other tools simply cause the results to be updated accordingly. The results panel is automatically populated. ii.) users can query for shots similar to one of the retrieved results by clicking the magnifying glass icon on the thumbnail. This causes an internal database query which is handled faster than QbS queries, since no feature extraction is required. This is *Query-by-Example* (QbE). iii.) by marking relevant (+) and irrelevant (-) shots in the results panel, a user can provide *relevance feedback*. The adjusted results will be closer to the relevant shots and further away from the irrelevant ones, in feature space.

Architecture

The architecture of IMOTION is shown in Figure 1. In the pre-processing (offline) phase, the collection is broken into shots using video segmentation. Resulting shots are processed by a set of visual feature extractors that index visual appearance i.e., color, edge, motion and a set of semantic extractors that label videos using entry-level categories e.g., "furniture" or "cow". Resulting feature vectors are inserted into the database.

In the online phase, the color, motion, and line sketches are separated based on the drawing tool used. The line sketch is periodically grossly segmented and submitted to a state-of-the-art sketch recognizer, which outputs at most three suggestions for the completion of the user's object (see Figure 2). If the user accepts a completion suggestion e.g., car, the chosen object category is mapped to one or more matching entry-level categories e.g., "sedan" and "vehicle" using a simple syntactic similarity matrix. As a result, only the shots containing cars are retrieved from the database.

For the color sketch and motion trajectories, the visual feature extractors are run in real-time to extract query feature vectors. The database backend retrieves optimally top-k nearest neighbors of each query vector. Distance-based scores for each retrieved shot undergo score fusion with weights specified by sliders in the UI. Results from the query are retrieved asynchronously and the list is continually updated to reflect order changes.

Features and Fusion

The features used by IMOTION can be grouped in four different categories; color, edge, motion, and semantic features. The results from these features are combined in a two-stage score-based late fusion approach. The first stage is performed by the back-end using statically optimized feature weights per category. This means that for every category only one list of result candidates needs to be transferred to the UI. The second combination step happens in the UI where the per-category results are combined in real-time into one result set based on user defined category weights. This also means that the result display is updated live whenever new information from the retrieval back-end becomes available thereby highly increasing the overall responsiveness of the system.

Figure 2: IMOTION QbS automcompletion in action: (a) user sketches a color sketch with the brush tool; matching images of the sea are retrieved (b) user switches to pencil and starts drawing a boat; a suggestion pop-up appears mid-sketch; (c) user selects the first option and the new results are displayed

The Demo
At the demo, conference participants will be able to use the IMOTION UI to search for video shots on the basis of sketches they draw. To allow for known item search, they will be able to browse through the video collection on a separate computer before drawing their query sketch.

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), the Scientific and Technological Research Council of Turkey (Tübitak, grant no. 113E325), and the Swiss National Science Foundation (SNSF, contract no. 20CH21_151571).

Sketch recognition and concept mapping
Sketch autocompletion is achieved in three stages. At any time the user interacts with the sketchboard via the pencil, the bounding box of the sketch is updated. The tracked boxes thus contain partial or completed sketches of different objects. The system passes these segmented sketches to a sketch recognizer capable of understanding partial sketches [12]. The sketch recognition system takes as input a binary (B/W) sketch and outputs probability scores for 250 objects (from the sketch dataset [4]). Top 3 probabilities are then presented as sketch completion suggestions to the user via a pop-up.
Upon clicking on a suggestion, the user's incomplete sketch is replaced by the full sketch he or she chose (see Figure 2(c)). In order to retrieve the selected object, the selected sketch class (e.g., "sailboat") must be converted to one or several (out of 325) semantic categories that index the collection discussed in the previous section. We achieve this via simple computational linguistics tools: we map the 250 sketch categories and the 325 semantic categories to "Synsets" using WordNet tools. We then compute a 250×325 similarity matrix based on Wu-Palmer distance [15]. This allows to select the semantic concepts closest to the sketched object. In the case of "sailboat", the three closest entry-level concepts are "sailboat" (a perfect match), "boat" and "speed-boat".

Database
IMOTION is based on the distributed storage system ADAM [5] that allows to easily manage, organise, and query multimedia objects, i.e., the corresponding structured metadata, as well as the extracted high-dimensional feature vectors. ADAM uses PostgreSQL and comes with the Vector Approximation-File [14] indexing and k nearest neighbor search. The evaluation of ADAM has shown that it is able to handle big multimedia collections of multiple million objects and keep the retrieval time well below a few seconds (e.g., for 14 million elements each storing 144 dimensions, ADAM returns results on average in 0.55 seconds for the 100 most similar objects [5]).

Video collection
For this demo, we use the Open Short Video Collection 1.0 [8] which consists of 200 creative commons licensed videos with a large visual diversity. Its roughly 20 hours of video are segmented into over 30k shots to enable retrieval with a high degree of temporal accuracy.

Related work
Content based image and video retrieval systems are routinely evaluated in annual challenges such as ILSVRC [10] and TRECVID [7]. The focus is on tagging a large collection with a high number of pre-determined classes and not on interactive search. ILSVRC15 evaluates 1,000 categories on a test dataset of 150,000 photographs (object localization challenge) and 401 scene categories on 381,000 images. In TRECVID 2014 Semantic Indexing, participants provide top candidate shots for 500 semantic concepts from a slice of the IACC.2 collection (7,300 videos, 600 hours of video in total). Interactive sketch symbol recognition has achieved impressive results in recent times. Trained on 20,000 crowdsourced sketches, the SVM-based classifier in [3] can recognize in real-time 250 object categories. Sketch autocompletion has been studied in [13, 2]. However when used for retrieval purposes, Sketch Based Image Retrieval (SBIR) and Sketch Based Video Retrieval (SBVR) seem relatively limited in scope. The largest SBIR dataset we found in the literature [6] is of 15,000 Flickr images, querried by 33 line sketch categories. Similarly, a SBVR system based on motion sketching[1] can perform retrieval in 1000+ video clips. Both these systems report retrieval time above 2 seconds.

REFERENCES

1. Chiranjoy Chattopadhyay and Sukhendu Das. 2012. A motion-sketch based video retrieval using MST-CSS representation. In *ISM 2012*. IEEE, 376–377.
2. Gennaro Costagliola, Mattia De Rosa, and Vittorio Fuccella. 2014. Recognition and autocompletion of partially drawn symbols by using polar histograms as spatial relation descriptors. *Computers & Graphics* 39 (2014), 101–116.
3. Mathias Eitz, James Hays, and Marc Alexa. 2012. How Do Humans Sketch Objects? *ACM Trans. Graph. (Proc. SIGGRAPH)* 31, 4 (2012), 44:1–44:10.
4. Mathias Eitz, Kristian Hildebrand, Tamy Boubekeur, and Marc Alexa. 2011. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *Visualization and Computer Graphics, IEEE Transactions on* 17, 11 (2011), 1624–1636.
5. Ivan Giangreco, Ihab Al Kabary, and Heiko Schuldt. 2014. ADAM - A Database and Information Retrieval System for Big Multimedia Collections. In *International Congress on Big Data*. IEEE, 406–413.
6. Rui Hu and John Collomosse. 2013. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *Computer Vision and Image Understanding* 117, 7 (2013), 790–806.
7. Paul Over, George Awad, and Martial Michel et al. 2015. TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. TRECVID 2015*. NIST, USA.
8. Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. 2015a. *OSVC – Open Short Video Collection 1.0*. Technical Report CS-2015-002. University of Basel.
9. Luca Rossetto, Ivan Giangreco, Heiko Schuldt, Stéphane Dupont, Omar Seddati, Metin Sezgin, and Yusuf Sahillioğlu. 2015b. IMOTION – A Content-Based Video Retrieval Engine. In *MultiMedia Modeling*. 255–260.
10. Olga Russakovsky, Jia Deng, and Hao Su et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.
11. Klaus Schoeffmann, David Ahlström, and Werner Bailer et al. 2014. The video browser showdown: a live evaluation of interactive video search tools. *International Journal of Multimedia Information Retrieval* 3, 2 (2014), 113–127.
12. Omar Seddati, Stephane Dupont, and Saïd Mahmoudi. 2015. DeepSketch: Deep convolutional neural networks for sketch recognition and similarity search. In *CBMI 2015*. IEEE, 1–6.
13. Caglar Tirkaz, Berrin Yanikoglu, and T Metin Sezgin. 2012. Sketched symbol recognition with auto-completion. *Pattern Recognition* 45, 11 (2012), 3926–3937.
14. Roger Weber, Hans-Jörg Schek, and Stephen Blott. 1998. A Quantitative Analysis and Performance Study for Similarity-Search Methods in High-Dimensional Spaces. In *VLDB 1998: International Conference on Very Large Data Bases*. New York, USA, 194–205.
15. Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*. 133–138.