IMOTION — A Content-Based Video Retrieval Engine

Luca Rossetto[1], Ivan Giangreco[1], Heiko Schuldt[1], Stéphane Dupont[2], Omar Seddati[2], Metin Sezgin[3], and Yusuf Sahillioğlu[3]

[1] Databases and Information Systems Research Group,
Department of Mathematics and Computer Science, University of Basel, Switzerland
{firstname.lastname}@unibas.ch

[2] Research Center in Information Technologies, Université de Mons, Belgium
{firstname.lastname}@umons.ac.be

[3] Intelligent User Interfaces Lab, Koç University, Turkey
{mtsezgin,ysahillioglu}@ku.edu.tr

Abstract. This paper introduces the IMOTION system, a sketch-based video retrieval engine supporting multiple query paradigms. For vector space retrieval, the IMOTION 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, the IMOTION 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.

1 Introduction

The IMOTION content-based video search engine is being developed in the context of the Chist-Era project IMOTION [2] (Intelligent Multi-Modal Augmented Video Motion Retrieval System), a joint effort of the Numediart Institute for Creative Technologies at the University of Mons (Belgium), the Intelligent User Interfaces Lab at Koç University (Turkey), and the Databases and Information Systems Research Group at the University of Basel (Switzerland). The IMOTION system is based on Cineast [7], which has originally been designed and implemented for sketch-based known-item video retrieval applications. It is capable of retrieving video sequences based on either single frames or rough color- or edge-sketches. However, it also supports additional retrieval modes like query-by-example, or motion queries.

The IMOTION system does that by combining multiple low-level features (e.g., color, edge, and motion features) with high-level spatial and temporal features (e.g., keyframe content, motion). The various features and the meta data are all stored in the database and information retrieval system ADAM [1], a storage engine built upon PostgreSQL, that jointly supports Boolean retrieval (exact search) and vector space retrieval (similarity search).

X. He et al. (Eds.): MMM 2015, Part II, LNCS 8936, pp. 255–260, 2015.

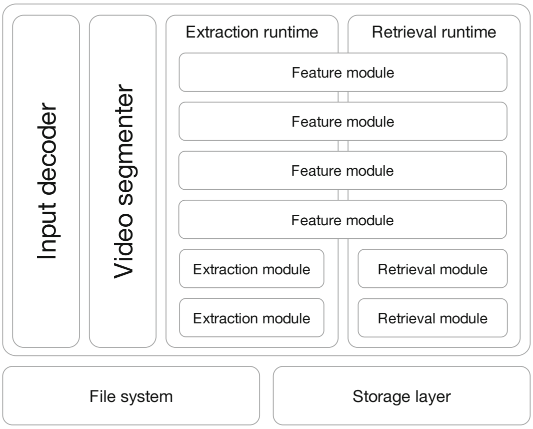


Fig. 1. Architectural overview of IMOTION

The user interface of the IMOTION system is browser-based and provides a sketching canvas for query specification. It has been optimized for use on a tablet computer, but can also be used on any other hardware.

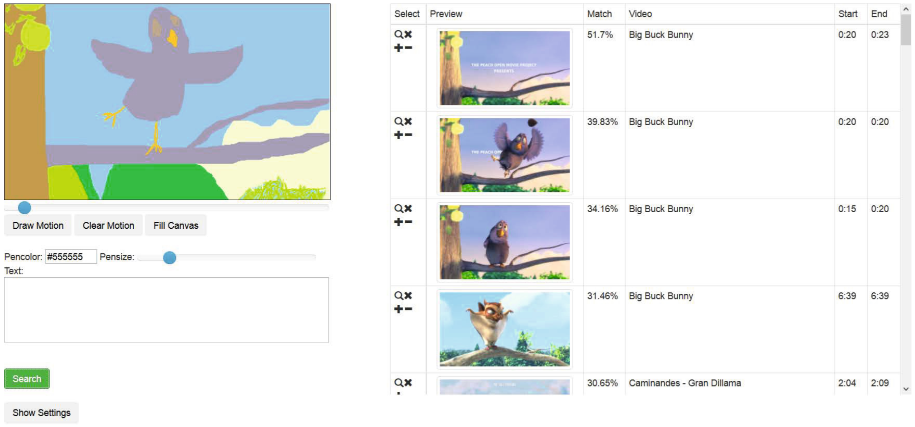
The remainder of this paper is structured as follows: Section 2 introduces the system architecture of IMOTION. In Section 3, we briefly present the features and in Section 4 the retrieval modes supported by the IMOTION system. Section 5 concludes.

2 IMOTION System Architecture

From a conceptional point of view, the IMOTION system can be divided into an off-line part, responsible for feature extraction and management, and an on-line part which handles the actual retrieval. Figure 1 illustrates the architecture of the IMOTION system.

The *on-line* part consists primarily of a module runtime which manages the individual feature modules of the features supported by IMOTION. These modules work in parallel and independently of each other. The module runtime handles the initialisation of the modules, provides them with the input information they need, manages their outputs and shuts them down when they are no longer needed. During retrieval, the module runtime receives a query object which it passes to the retrieval modules. The modules, having been initialised with a connection to the storage layer, query the storage engine for shots matching the query object. Each module independently returns result information which is finally combined by the runtime to determine the overall result set of a query.

The *off-line* part consists of an input decoder, a video segmenter and the individual extraction modules for the IMOTION features. Input decoding logic provides the video segmenter with a continuous stream of data which it segments into shots. These shots are then passed to the feature extraction modules,



Pencolor: #555555 Pensize:
Text:

Select	Preview	Match	Video	Start	End
🔍✖ +−		51.7%	Big Buck Bunny	0:20	0:23
🔍✖ +−		39.83%	Big Buck Bunny	0:20	0:20
🔍✖ +−		34.16%	Big Buck Bunny	0:15	0:20
🔍✖ +−		31.46%	Big Buck Bunny	6:39	6:39
🔍✖		30.65%	Caminandes - Gran Dillama	2:04	2:09

Fig. 2. Web-based user interface of the IMOTION system

which then perform the extraction of the corresponding features which are finally handed off to the storage layer.

Usually, a module performs both the retrieval and extraction task because the logic for transforming video information into feature representation is the same in both phases. An exception to this are those retrieval modules which use data already generated by other extraction modules.

The IMOTION system provides a browser-based web-interface offering a painting canvas where existing images can be pasted and possibly also manipulated and where sketches can be provided (either edge sketches or color sketches). Furthermore, it also allows to specify flow-fields for the specification of motion. The results of a search are presented in a list which contains a representative keyframe of the sequence as well as additional meta information. Figure 2 shows an example of a query and its results in the browser-based web-interface.

3 Image and Motion Features

The IMOTION system makes use of a multitude of features which are described in more detail in the following sections.

3.1 Low-Level Features

The set of low-level features contains features concerned with color and edges (representing spatial information) and motion (representing temporal information). The color features describe global as well as regional color properties of a shot, such as average or median color of a specific region. The edge features consider the regional distribution as well as directionality of edges. Finally, the motion features produce regional histograms of directions of movement. A complete list of the low-level features implemented in the IMOTION system is presented below.

All low-level features use the quadratic Euclidean distance for comparison of feature vectors. The distances are converted into a similarity score using a linear transformation for which the maximum is determined empirically. The results of the single feature modules are combined to a single coherent result set by computing a weighted average over the similarity scores.

A complete list of the low-level features implemented in the IMOTION system is presented below.

Global features

- *Average / Median color*
- *Dominant shot colors*: centre-points of the three largest color clusters
- *Chroma / Saturation*: average of all chroma / saturation values of a shot
- *Color histogram*
- *Shot position*: relative position of a shot with respect to the entire video

Regional color features

- *Color moments*: channel-wise statistical moments over regional partitions (uniform grid, angular radial partitioning) of an aggregation over all frames of a shot
- *Registered color grid*: grid of fuzzy quantized colors registered during retrieval.
- *Color layout descriptor* [3]
- *Color element grids*: grids containing partial color information in various representation (average saturation, variance of hue, etc.)
- *Subdivided color histogram*: fuzzy color histograms of image partitions

Regional edge features

- *Partitioned edge image*: regional ratios of edge- and non-edge pixels
- *Edge histogram descriptor* [6]
- *Dominant edge grid*: regional dominant edge direction quantised into 5 categories.

Motion features

- *Directional motion histograms*: regional normalised histograms of motion quantized into 8 directions.
- *Regional motion sums*: regional sums of the lengths of all motion vectors.

3.2 High-Level Features

The category of high-level features makes use of state-of-the-art machine learning approaches based on deep neural networks to extract relevant descriptors. Here too, two categories of features are available, representing either spatial (keyframe appearance), or temporal (video shot motion) information. Such approaches are able to efficiently encode natural image (and motion) key characteristics and similarities.

For the spatial component, we use a neural network architecture similar to the one proposed by Krizhevsky et al. [4]. The training is conducted on the image dataset ImageNet [8], which contains 1000 categories and about 1.2 million images. The available diversity and natural characteristics in the training data are very important to reach to a system better able to generalize to unseen image content. For the spatial component, dedicated to extracting motion-related features, we also use a similar architecture but rather than image pixels, the input relies on optical flow extracted from the video shots, and which direction components are considered as if they were image channels. Training is performed on the following video data sets: KTH [9], UCF101 [10], HMDB-51 [5].

Multiple techniques are applied to calculate the optical flow locally in time, and a sequence of these micro-movements (representing motion within the video shot) is passed as input to the neural network. The outputs of the last hidden layer of these two neural networks are then used as high-level features amenable to the vector space retrieval approach used in the IMOTION system.

4 Retrieval Modes

The IMOTION system supports known-item search by offering users three different modes of retrieval. The first mode, *query-by-sketch*, is a direct user input mode in which the user provides a rough hand-drawn sketch (either a line drawing or a color sketch) to search within the video collection. The second mode offers a *query-by-example* interaction where a user provides a query object via drag-and-drop.

This approach can be used when a user wants to find sequences similar to a previously retrieved one. It is achieved by using the internal representation of a shot as input and otherwise proceeding as usual. The third mode, *motion queries*, allows a user to specify the motion of objects across consecutive frames via (partial) flow-fields. These three retrieval modes are complemented with *relevance feedback* for refining the query results (i.e., by marking either relevant or irrelevant elements from the result list of a previous search).

The resulting query will then produce results which are similar to the relevant set, but not to the non-relevant set. Note that IMOTION has a stateless behavior and does not employ learning methods; feedback from previous iterations is, thus, discarded and no session is kept for queries.

5 Conclusion

In this paper, we have presented the IMOTION system, a content-based video retrieval engine for known-item searches using exemplary images or sketches. Since the IMOTION system was developped to support a wide variety of different kinds of video and implements many diverse features (both low-level and high level) and query paradigms that can be flexibly combined, it provides effective support for known-item search in large video collections.

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).

References

1. Giangreco, I., Kabary, I.A., Schuldt, H.: ADAM — A Database and Information Retrieval System for Big Multimedia Collections. In: Proc. Int. Congr. on Big Data 2014 (BigData 2014), Anchorage, USA. IEEE (2014)
2. IMOTION project, `https://imotion-project.eu/`
3. Kasutani, E., Yamada, A.: The MPEG-7 Color Layout Descriptor: A Compact Image Feature Description for High-Speed Image/Video Segment Retrieval. In: Proc. Int. Conf. on Image Processing (ICIP 2001), Thessaloniki, Greece, pp. 674–677. IEEE (2001)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 25: Proc. Conf. on Neural Information Processing Systems (NIPS 2012), Lake Tahoe, USA, pp. 1097–1105 (2012)
5. Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T.: HMDB: A Large Video Database for Human Motion Recognition. In: Proc. Int. Conf. on Computer Vision (ICCV 2011), Barcelona, Spain, pp. 2556–2563. IEEE (2011)
6. Park, D.K., Jeon, Y.S., Won, C.S.: Efficient Use of Local Edge Histogram Descriptor. In: Proc. Ws. on Multimedia, Los Angeles, USA, pp. 51–54. ACM (2000)
7. Rossetto, L., Giangreco, I., Schuldt, H.: Cineast: A Multi-Feature Sketch-Based Video Retrieval Engine. In: Proc. Int. Symp. on Multimedia (ISM 2014), Taichung, Taiwan. IEEE (December 2014)
8. Russakovsky, O., Deng, J., Su, H., et al.: ImageNet Large Scale Visual Recognition Challenge. CoRR, abs/1409.0575 (2014)
9. Schüldt, C., Laptev, I., Caputo, B.: Recognizing Human Actions: A Local SVM Approach. In: Proc. Int. Conf. on Pattern Recognition (ICPR 2004), Cambridge, England, pp. 32–36. IEEE (2004)
10. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. CoRR, abs/1212.0402 (2012)