Finding the Chameleon in Your Video Collection

Marco A. Hudelist
Klagenfurt University,
Klagenfurt, Austria
marco@itec.aau.at

Christian Beecks
RWTH Aachen University,
Aachen, Germany
beecks@informatik.rwth-aachen.de

Klaus Schoeffmann
Klagenfurt University,
Klagenfurt, Austria
ks@itec.aau.at

ABSTRACT

We present a novel content-based video retrieval tool that facilitates interactive search in large video archives by focusing on the factors content context and content dynamics. It incorporates query-by-concept and query-by-sketch functionality. For the latter we introduce temporal feature signatures - an extension to color feature signatures by adding the dimension of content dynamics over time. Moreover, temporal feature signatures are also used for performing segment similarity searches for improved matching performance in contrast to static solutions. Because of intelligent caching this process is performed in just a couple of seconds, which improves the overall user experience greatly. Found segments can also be easily displayed in their chronological context for refining the search. To better visualize the content dynamics of a video segment users just need to move their mouse cursor across a segments thumbnail to peek into its content.

CCS Concepts

•Information systems → Multimedia information systems; •Human-centered computing → Ubiquitous and mobile computing systems and tools; •Computing methodologies → Visual content-based indexing and retrieval; Object identification;

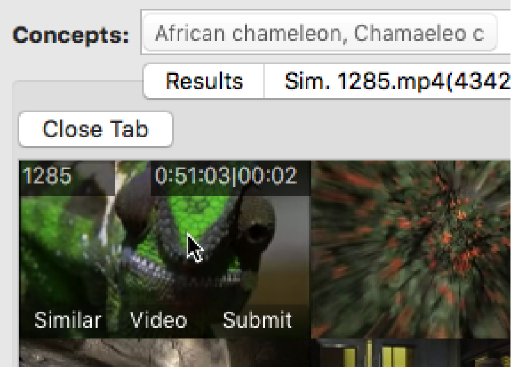
Keywords

Video retrieval; feature signatures; collaborative search; human computer interaction

1. INTRODUCTION

There are many proposals for improving content-based video retrieval. The majority of video retrieval tools use query-by-text, query-by-example, query-by-sketch or a combination of those approaches. Although these methods are technically sound users see themselves often frustrated in using such tools. The produced result lists are most of the time not a perfect match for what users were actually looking for. Contrary to the belief that a thoughtful query gets users directly to the right result this is often not the case. The reasons for this are the *semantic gap* [15, 4], limitations in terms of content analysis accuracy, or the *usability gap* [14]. Furthermore, users often have a hard time transforming their mental image of what they are looking for to a query language that a retrieval system can process. In their mind they can only create an "abstract query" - a rough description of the scene with usually few but very specific details. It is often not possible for users to translate this abstract query to the features supported by the video retrieval tool. Another frustration occurs when users try to refine the initial results. Most content-based video retrieval (CBVR) tools have only poor support for this phase of search. Their browsing interaction design is lacking, which is problematic, as it is essential for the search process. For example, many tools do not allow users to easily go back to an earlier result after they have submitted a new query. Moreover, they ignore the dynamic aspects of video content as everything is represented by static thumbnails. However, it is often only possible to distinguish very similar segments by understanding their dynamic content, e.g. let them play. It is also important to show a segment's position in its parent video for situations where users realize that the wanted segment is surely in the same video.

In contrast, even simple but well-designed interfaces can outperform sophisticated video retrieval systems, as has been shown in the Video Browser Showdown (VBS) competition



Concepts: African chameleon, Chamaeleo c

Figure 1: Zoomed view of a segment thumbnail after placing the mouse above it. Top: video name, start time and duration. Bottom: buttons for similarity search and temporal segment browsing.

MMSys'16 May 10-13, 2016, Klagenfurt, Austria

ACM ISBN 978-1-4503-4297-1/16/05.
DOI: http://dx.doi.org/10.1145/2910017.2910631

[13] of 2015 by Huerst et al. [9] and in 2012 by Del Fabro and Böszörményi [7]. Simple and human-computation-based approaches have their limits though, especially in terms of data set sizes and user fatigue. A good search tool has to find a balance between query-based retrieval and interactive human-based result refining.

In this work we want to address this issue and present a tool for fast interactive video search and filtering in large video archives. It offers users two options for defining queries: (i) query-by-concept by choosing from a list of available concepts, and (ii) sketching by defining a temporal feature signature of the wanted segment. Temporal feature signatures are an extension of color feature signatures as shown by Beecks et al. [2]. They incorporate the dynamic nature of video segments to improve results for sketching and segment similarity searches. For the refinement phase, the system enables users to start a segment similarity search for each segment - again based on temporal feature signatures - or display a segment within its parent video in its chronological context. Moreover, it features dynamic interactive thumbnails for easy and fast content inspection and supports thread-based search approaches (also called aspect-based search [5, 16]) to let users explore multiple search strategies at the same time.

The presented tool is an extension of the system used for the VBS 2016 competition [8]. Our "Collabris"-system additionally incorporates collaborative aspects by exchanging inspection data with a tablet tool that focuses on fast human-based inspection. The system scored exceptionally well, especially in the very challenging textual search session. Overall, we achieved the second place in the expert sessions and third in the novice sessions with this tool.

2. MEDIA ANALYSIS

For content analysis the system uses an offline/online approach. In the offline phase the required metadata for concept search and sketch search is generated. The pre-generated metadata is then used in the online phase to guarantee a smooth user experience.

2.1 Shot Detection With Optical Flow

In the first step we divide each video into a number of shots, based on optical flow. For that purpose we start with an initial set of densely sampled points in the frame and track them with the Kanade-Lucas-Tomasi (KLT) algorithm [3] from one frame to another. As soon as the number of track-able points falls below a specific threshold t_C we detect a shot change and restart the tracking with a fresh set of densely sampled points. For each detected shot the middle frame is selected as keyframe.

2.2 Temporal Feature Signatures

Temporal feature signatures advance the feature signature model by taking temporal characteristics of features into account. In particular, they facilitate dynamic shot-wise content aggregation by utilizing object-specific feature quantizations.

For each video shot, we first extract the characteristic key frames and model the content-based properties of each single key frame by means of features $f_1, \ldots, f_n \in \mathbb{F}$ in a feature space $\mathbb{F}$. In order to reflect the perceived visual properties of the frames, we utilize a 7-dimensional feature space $\mathbb{F} = \mathbb{R}^7$ comprising spatial information, CIELAB color information [10], coarseness, and contrast information. By clustering the extracted local feature descriptors with the k-means algorithm, we obtain a feature signature $S : \mathbb{F} \to \mathbb{R}$ subject to $|\{f \in \mathbb{F} | S(f) \neq 0\}| < \infty$ for each single key frame, where the representatives $\mathrm{R}_S = \{f \in \mathbb{F} | S(f) \neq 0\} \subseteq \mathbb{F}$ are determined by the cluster centroids and their weights $S(f)$ by the relative frequencies of the cluster centroids (for further details see Beecks [1]).

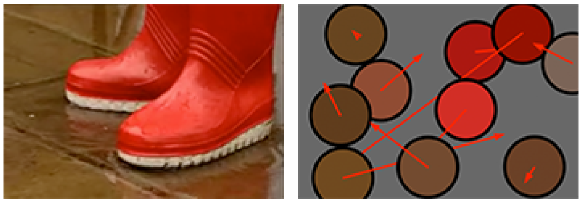


Figure 2: Sample visualization of an analysis result of temporal feature signatures.

Based on this adaptive-binning feature representation model, the spatial change of the cluster centroids over time within a single shot is taken into account. To this end, each video shot is modeled by a temporal feature signature $\widetilde{S} \in \mathbb{R}^{\widetilde{\mathbb{F}}}$ which extends the feature signature of the video shot's first key frame by tracking the spatial movement of the cluster centroids. By assigning each cluster centroid from the first frame to its nearest counterpart in the next frame based on the Euclidean distance and repeating this assignment until the last frame of a video shot is reached, the resulting spatial position of each cluster centroid are obtained. This spatial position is stored in two additional dimensions of the extended feature space $\widetilde{\mathbb{F}} = \mathbb{R}^9$ and hence defines the temporal feature signature $\widetilde{S}$ (see Fig. 2 for an example).

Based on the temporal feature signatures described above, an asymmetric variant of the Signature Matching Distance [2] is utilized in order to efficiently compare two video shots with each other. Given two temporal feature signatures $\widetilde{S}_x, \widetilde{S}_y \in \mathbb{R}^{\widetilde{\mathbb{F}}}$, their dissimilarity is defined as follows:

$$\mathrm{D}_\delta(\widetilde{S}_x, \widetilde{S}_y) = \sum_{(f,g) \in \mathrm{m}^{\delta\text{-NN}}_{\widetilde{S}_x \to \widetilde{S}_y}} \widetilde{S}_x(f) \cdot \delta(f, g),$$

where $\mathrm{m}^{\delta\text{-NN}}_{\widetilde{S}_x \to \widetilde{S}_y}$ is the nearest neighbor matching that relates similar features to each other based on a ground distance $\delta : \widetilde{\mathbb{F}} \times \widetilde{\mathbb{F}} \to \mathbb{R}$ that models the dissimilarity between two individual features. We utilize the Manhattan distance as ground distance, as this shows higher performance in terms of both efficiency and accuracy than the Euclidean distance.

2.3 Concept Detection

Concept-based filtering is supported by selection of visual classes that were trained on ImageNet [6]. For this, we employ deep learning with convolutional neural networks (CNN), utilizing the freely available Caffe framework [11]. Moreover, we use the "*BVLC AlexNet*" model trained on ILSVRC 2012 data [12], which is freely available on the website of Caffe [11]. With that model we classify each keyframe of a shot and use the five concepts with highest confidence as a result. Our tool provides a filter function based on the confidence value; i.e., the user can filter for detected concepts with a specific minimum confidence only.

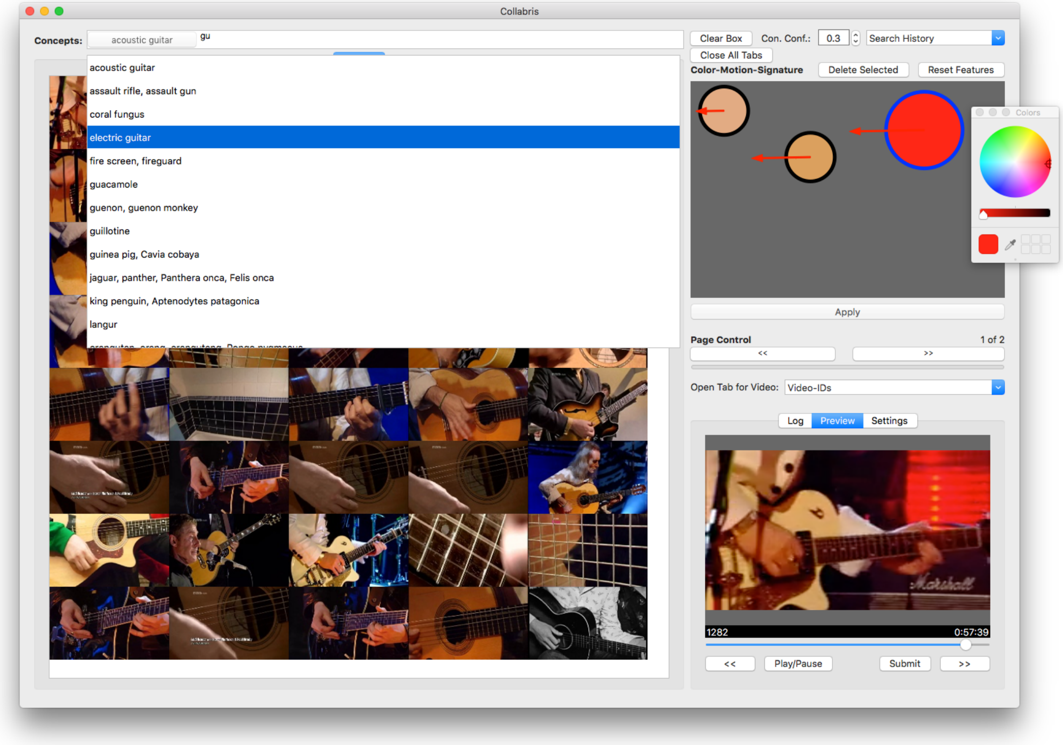


Figure 3: CBVR tool with controls for sketching temporal feature signatures (top right), concept filtering (top), preview player (bottom right), tabs for chronological segment browsing (top, behind concept suggestions box) and search history (top right).

3. INTERFACE

The interface offers various options for filtering and sorting of video segments. On one hand, users can activate filters for specific concepts by typing in the concept box at the top (see Fig. 3). A pop-up shows matching concept names that are available. It is also possible to combine multiple concepts. Users have the option to set the required minimum concept confidence level right next to the concept box.

Another way to resort video segments is by using a temporal feature signatures sketch. In Fig 3 the sketching area is located at the top right, visualized by a large gray drawing area - an abstract representation of a frame. Color clusters can be added to the sketch by simply clicking on the sketch area. The new cluster is then added at the clicked location and a color panel appears for further configuration. Moreover, it is possible to change the size of the cluster by clicking and dragging a clusters' outer border. As temporal feature signatures also support setting a direction in which a color cluster is expected to move over time this property can be set as well. Users just have to right-click and drag their mouse cursor over a selected cluster and therefore define the direction. Visually this is indicated by a red arrow attached to the cluster (see Fig 3). Results of an applied filtering process - either concept-based, signature-based or both - are displayed in the middle of the interface. Each result segment is represented by a keyframe extracted from the middle of the segment. If there are more result segments as can be currently displayed, additional result pages can be scrolled by using paging controls, which are located right blow the sketch area.

To understand the dynamic context, users simply have to place their mouse cursor on a keyframe. The relative cursor position is then temporally mapped to the segment's content and the appropriate frame is displayed. Therefore, users just need to move their mouse cursor from left to right to inspect the whole segment. Moreover, various segment related information is displayed when users hover over it as can be seen in Fig. 1. At the top the video id is displayed as well as the start time code of the segment and its duration. At the bottom buttons are available to start a search for visually similar segments based on their temporal feature signatures and open a new tab to display the segment in the chronological context of its parent video.

It is also possible to play a video segment by clicking on it. At the bottom right of the interface a preview player then automatically starts playback of the segments contents. Furthermore, the player lets users inspect the related video outside the segments' boundaries via playback controls.

All sorting and filtering actions and their result are stored for quickly returning to an earlier state. This "search history" is accessible through a combo box at the top right of the interface. The entries are chronologically ordered so that the latest search configuration is always at the top. Choosing an entry immediately updates the interface accordingly. Moreover, users can open a new tab directly to browse a videos' segments chronologically by utilizing a combo box located beneath the paging controls. It provides the video IDs of all available videos and choosing any entry opens a tab and shows the first segments of the related video.

4. WORKFLOW EXAMPLE

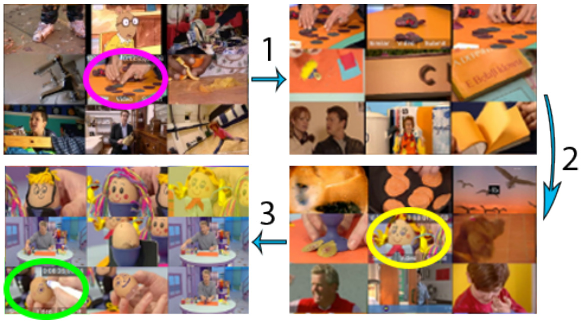


Figure 4: Workflow of searching a scene of a *"man painting a smiley on an egg"*: (1) similarity search, (2) scrolling through results, (3) temporal segment browsing.

A typical workflow can be described as follows. A scene with a *"man painting a smiley face on an egg"* with a close-up of his hands should be found. To start the search users might want to think about what to expect in the scene. It may has something to do with handicraft and typically you need a plane for working. Therefore, users select the concept *"woodworking plane"* from the list, as it comes closest to what they have in mind. The returned results are good but not exactly what users were looking for. Nevertheless, they recognize a segment with a close-up of the hands working on a plane and start a similarity search. In the returned results they find a scene with an egg and a close-up of hands, but the painted smiley is already finished. Therefore, they open the chronological view for this segment, perform a little bit of scrolling and quickly find the desired scene. See Fig. 4 for a simplified visualization of this workflow.

5. CONCLUSIONS

In this demo paper we presented a content-based video retrieval tool for efficient search in large video archives. It supports filtering for semantic visual concepts trained with convolutional neural networks on ImageNet as well as query-by-sketch searches based on our novel temporal feature signatures approach. To support users in refining their initial search results the interface makes it easy to perform similarity searches for segments or temporal browsing. The system was successfully tested in the VBS 2016 competition and placed second in the expert session. In future work we plan on evaluating our system in a user study to compliment our competition results.

6. ACKNOWLEDGMENTS

This work was supported by Universität Klagenfurt and Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF 20214 u. 3520/ 26336/38165.

7. REFERENCES

[1] C. Beecks. *Distance-based similarity models for content-based multimedia retrieval.* PhD thesis, RWTH Aachen University, 2013.

[2] C. Beecks, S. Kirchhoff, and T. Seidl. Signature matching distance for content-based image retrieval. In *ICMR*, pages 41–48, 2013.

[3] J.-Y. Bouguet. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel Corporation*, 5(1-10):4, 2001.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):5:1–5:60, May 2008.

[5] O. de Rooij, C. G. M. Snoek, and M. Worring. Query on demand video browsing. In *ACM Int. Conf. on Multimedia*, MM '07, pages 811–814, New York, NY, USA, 2007. ACM.

[6] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.

[7] M. Fabro and L. Böszörmenyi. *Multimedia Modeling*, chapter AAU Video Browser: Non-Sequential Hierarchical Video Browsing without Content Analysis, pages 639–641. Springer, Berlin, Heidelberg, 2012.

[8] M. A. Hudelist, C. Cobârzan, C. Beecks, R. Werken, S. Kletz, W. Hürst, and K. Schoeffmann. *MultiMedia Modeling*, chapter Collaborative Video Search Combining Video Retrieval with Human-Based Visual Inspection, pages 400–405. Springer, Cham, 2016.

[9] W. Hürst and R. van de Werken. Human-based video browsing - investigating interface design for fast video browsing. In *IEEE ISM 2015 (to appear)*. 2015.

[10] ISO. Iso 11664-4:2008 (cie s 014-4/e:2007) - colorimetry – part 4: Cie 1976 l*a*b* colour space @ONLINE, Mar. 2016.

[11] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proc. of the ACM Int. Conf. on Multimedia*, MM '14, pages 675–678, New York, NY, USA, 2014. ACM.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc., 2012.

[13] K. Schoeffmann. A user-centric media retrieval competition: The video browser showdown 2012-2014. *MultiMedia, IEEE*, 21(4):8–13, Oct 2014.

[14] K. Schoeffmann and F. Hopfgartner. Interactive video search. In *ACM Int. Conf. on Multimedia*, MM '15, pages 1321–1322, New York, NY, USA, 2015. ACM.

[15] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, Dec 2000.

[16] T. Urruty, F. Hopfgartner, D. Hannah, D. Elliott, and J. M. Jose. Supporting aspect-based video browsing: Analysis of a user study. In *ACM Int. Conf. on Image and Video Retrieval*, CIVR '09, pages 47:1–47:8, New York, NY, USA, 2009. ACM.