Collaborative Video Search Combining Video Retrieval with Human-Based Visual Inspection

Marco A. Hudelist[1(✉)], Claudiu Cobârzan[1], Christian Beecks[3], Rob van de Werken[2], Sabrina Kletz[1], Wolfgang Hürst[2], and Klaus Schoeffmann[1]

[1] Klagenfurt University, Universitätsstrasse 65-67, 9020 Klagenfurt, Austria
{marco,claudiu,ks}@itec.aau.at, skletz@edu.aau.at
[2] Utrecht University, PO Box 80.089, 3508TB Utrecht, The Netherlands
{vandewerken,huerst}@cs.uu.nl
[3] RWTH Aachen University, Aachen, Germany
beecks@cs.rwth-aachen.de

Abstract. We propose a novel video browsing approach that aims at optimally integrating traditional, machine-based retrieval methods with an interface design optimized for human browsing performance. Advanced video retrieval and filtering (e.g., via color and motion signatures, and visual concepts) on a desktop is combined with a storyboard-based interface design on a tablet optimized for quick, brute-force visual inspection. Both modules run independently but exchange information to significantly minimize the data for visual inspection and compensate mistakes made by the search algorithms.

Keywords: Video retrieval · Interactive search · Interaction design · Feature signatures

1 Introduction and Related Work

We introduce an approach that is inspired by two successful systems from last year's edition of the Video Search Showcase (VSS) [16,17] as well as our earlier work on mobile video browsers [8,9]. Blaźek et al. showed a system based on signature-based similarity models that won the contest. Huerst et al. [13] provided an interface for efficient visual inspection of the whole database. Although the latter has demonstrated that such a pure human-based approach can compete quite well with traditional retrieval systems for surprisingly large databases [12], there is a natural limit to how many files can be inspected in a given time. Our new approach therefore aims at combining the best of both worlds. A desktop-based video retrieval tool searches for relevant files using traditional querying and filtering. The results are used to change the order in which videos are inspected on a tablet-based visualization of the data. Moreover, the results of the human inspection are in turn considered in future filtering iterations.

The collaborative aspects of our work are inspired by [6,7], both using tablets for collaborative search in single videos and archives, respectively. A server

© Springer International Publishing Switzerland 2016
Q. Tian et al. (Eds.): MMM 2016, Part II, LNCS 9517, pp. 400–405, 2016.
DOI: 10.1007/978-3-319-27674-8_40

storing preprocessed information about archived videos is queried by multiple clients on tablets using content-related criteria. The clients share information like already inspected frames, and submitted queries. Furthermore, our work incorporates signature-based similarity models, which adapt to individual multimedia contents by using an adaptive feature quantization. They have been utilized in many different domains ranging from multimedia data [4,5,19] to scientific data [2].

2 Proposed Approach

The main idea of our approach is to combine human-based visual inspection with content-based filtering in order to facilitate quick navigational access into large video archives. The overall system architecture consists of a desktop-based retrieval tool and a tablet app for visual browsing that exchange information continuously.

2.1 Content-Based Video Retrieval (CBVR) Tool

The main objective of the desktop-based CBVR tool is to retrieve video segments based on content characteristics, such as visual features and semantic visual concepts. For the visual features we utilize a signature-based similarity model relying on the Signature Matching Distance [3], while for the visual concepts we use convolutional neural networks (CNNs) [15].

Signature-Based Similarity Model. Given a video segment, 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, 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]). Based on this adaptive-binning feature representation model, we propose to utilize the Signature Matching Distance as distance-based similarity measure due to its superior retrieval performance [3]. The Signature Matching Distance defines a distance value between two feature signatures $X, Y \in \mathbb{R}^{\mathbb{F}}$ by making use of a matching $\mathrm{m} \subseteq \mathbb{F} \times \mathbb{F}$ that relates similar features to each other, a cost function $\mathrm{c} : 2^{\mathbb{F} \times \mathbb{F}} \to \mathbb{R}$ that determines the dissimilarity of a matching, and a ground distance $\delta : \mathbb{F} \times \mathbb{F} \to \mathbb{R}$ that models the dissimilarity between two features as $\mathrm{SMD}_\delta(X, Y) = \mathrm{c}(\mathrm{m}_{X \to Y}) + \mathrm{c}(\mathrm{m}_{Y \to X}) - 2\lambda \cdot \mathrm{c}(\mathrm{m}_{X \leftrightarrow Y})$ for $0 \leq \lambda \leq 1$. The parameter λ models the exclusion of bidirectional matches and is used to parameterize the similarity model accordingly.

Concept Search Based on Deep Learning. Our video retrieval application also supports concept-based search according to visual classes trained on ImageNet with convolutional neural networks (CNNs) [18]. For that purpose, we use a pre-trained model on the ImageNet "ILSVRC-2012" dataset [15], available in the Caffe framework [14]. Each key frame is classified according to this approach and detected ImageNet classes are stored for the corresponding segment if their confidence value is above a given threshold (0.5 in our case). From the entire set of visual concepts (up to 1,000 different ones) only the detected ones are made available for search in the interface of the desktop-based CVBR tool.

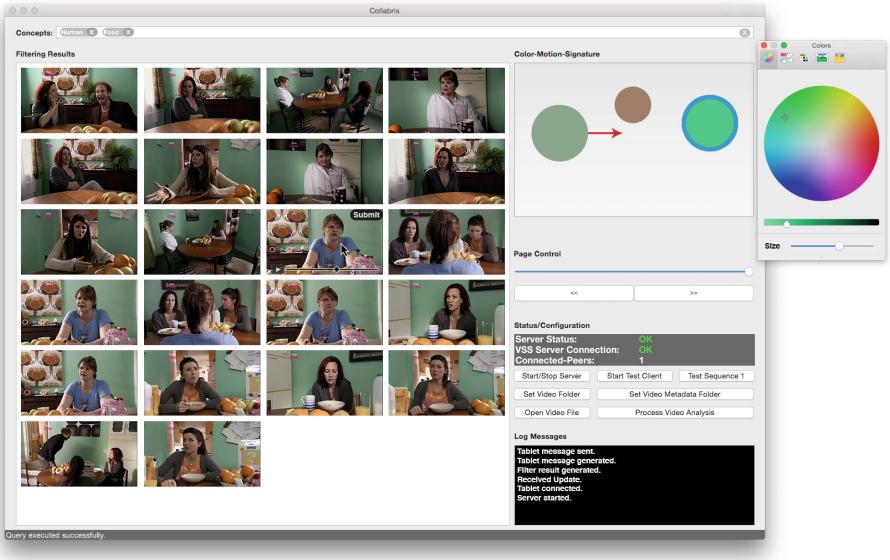

Fig. 1. Screenshot of the CBVR tool with display of color signature selector, concept textbox and preview options (Color figure online).

Interface. To define feature signatures users can utilize a blank canvas on the right hand side of the interface (see Fig. 1). The canvas represents a typical video frame. Location and color of a feature can be set by clicking on the canvas. This opens a color and size picker. Moreover, it is possible to define motion of the color feature by pressing and then dragging the mouse cursor in any direction. Motion is visualized with a red arrow. To define concepts users just have to start typing in the concept textbox at the top of the interface. Appropriate concepts start to appear automatically. Already selected concepts are displayed directly in the textbox as labels. Changes made in the filtering controls immediately update the result list. To further investigate a video segment in the result list users can use overlayed playback controls. Furthermore, a segments' screenshot is updated automatically when users hover over it, using the cursors x-coordinate for a temporal mapping. Moreover, a submit button is displayed as soon as users hover over a segment.

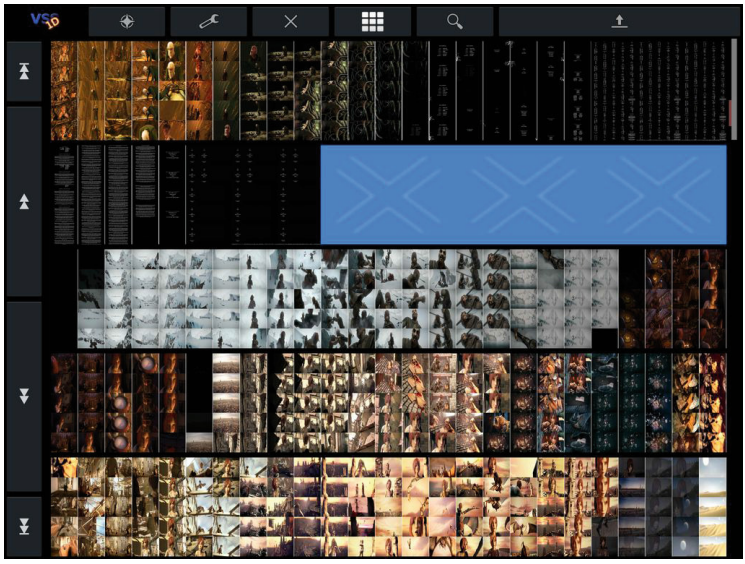

Fig. 2. Interface of the tablet app for human-based visual inspection.

2.2 Tablet App

The tablet app aims at enabling fast sequential search within the video archive by using an approach that proved successful in the VSS 2015. It employs a storyboard layout to present the whole video archive as a series of temporal arranged thumbnail images. The thumbnails are uniformly sampled every second from all the files within the archive. The interface (see Fig. 2) displays 625 images on one screen in accordance with previous research on optimal image sizes on mobiles [10,11]. The thumbnails are arranged in a mix of up/down-left/right directions in order to better identify scenes. The interaction options are kept to a minimum by employing only *Up* and *Down* actions; either across single screens or across files.

2.3 Collaboration Mechanism

All of the interaction between the two modules is performed automatically in the background. On one hand, the CVBR tool informs the tablet app about potentially promising videos, on the other hand, the tablet app informs the desktop-based CVBR tool about already inspected videos (Fig. 3). When the retrieval application generates a new segment result list it also generates a ranked list of videos for the tablet app. This list is created as follows: first, the original result list of segments is restricted to the top 250 results. Outgoing from this smaller list it is then counted how many segments are included in each video. The videos are then sorted accordingly to create a new ranked list for the tablet app. Videos with no matching segments are excluded from this list. The list

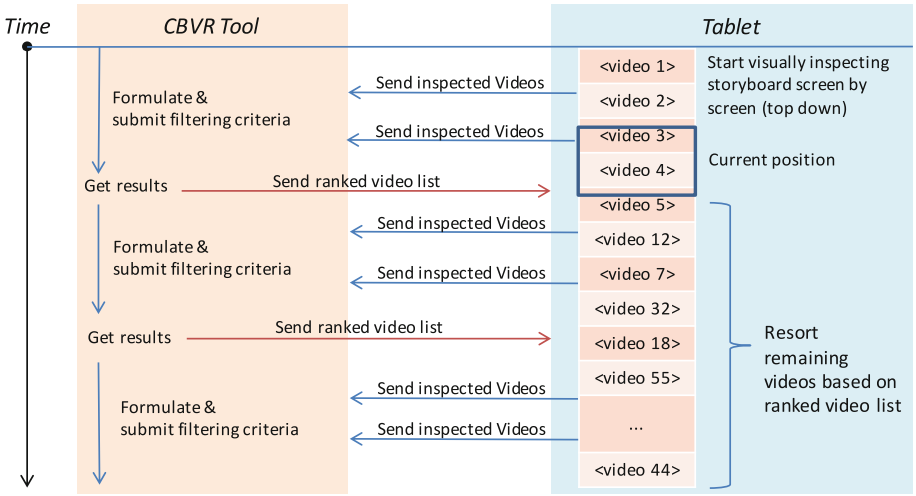

Fig. 3. Diagram explaining the collaboration between CBVR tool and tablet app.

is transmitted to the tablet app, which rearranges the remaining and not yet inspected videos accordingly. Furthermore, the tablet app gives continuous feedback to the desktop-based CVBR tool, about which videos have already been inspected by the tablet user. The CVBR tool then rearranges all corresponding video segments to the bottom of its result lists.

Acknowledgments. The work was funded by the Federal Ministry for Transport, Innovation and Technology (bmvit) and Austrian Science Fund (FWF): TRP 273-N15, supported by Lakeside Labs GmbH, Klagenfurt, Austria and funded by the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant 20214/26336/38165.

References

1. Beecks, C.: Distance-based similarity models for content-based multimedia retrieval. Ph.D. thesis, RWTH Aachen University (2013)
2. Beecks, C., Hassani, M., Hinnell, J., Schüller, D., Brenger, B., Mittelberg, I., Seidl, T.: Spatiotemporal similarity search in 3D motion capture gesture streams. In: Claramunt, C., Schneider, M., Wong, R.C.-W., Xiong, L., Loh, W.-K., Shahabi, C., Li, K.-J. (eds.) SSTD 2015. LNCS, vol. 9239, pp. 355–372. Springer, Heidelberg (2015)
3. Beecks, C., Kirchhoff, S., Seidl, T.: Signature matching distance for content-based image retrieval. In: ICMR, pp. 41–48 (2013)
4. Beecks, C., Kirchhoff, S., Seidl, T.: On stability of signature-based similarity measures for content-based image retrieval. MTAP **71**(1), 349–362 (2014)
5. Blažek, A., Lokoč, J., Matzner, F., Skopal, T.: Enhanced signature-based video browser. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015, Part II. LNCS, vol. 8936, pp. 243–248. Springer, Heidelberg (2015)

6. Cobârzan, C., Del Fabro, M., Schoeffmann, K.: Collaborative browsing and search in video archives with mobile clients. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015, Part II. LNCS, vol. 8936, pp. 266–271. Springer, Heidelberg (2015)
7. Cobârzan, C., Hudelist, M.A., Del Fabro, M.: Content-based video browsing with collaborating mobile clients. In: Gurrin, C., Hopfgartner, F., Hurst, W., Johansen, H., Lee, H., O'Connor, N. (eds.) MMM 2014, Part II. LNCS, vol. 8326, pp. 402–406. Springer, Heidelberg (2014)
8. Hudelist, M.A., Schoeffmann, K., Boeszoermenyi, L.: Mobile video browsing with the thumbbrowser. In: Proceedings of the 21st ACM International Conference on Multimedia, MM 2013, pp. 405–406. ACM, New York (2013)
9. Hudelist, M.A., Schoeffmann, K., Xu, Q.: Improving interactive known-item search in video with the keyframe navigation tree. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015, Part I. LNCS, vol. 8935, pp. 306–317. Springer, Heidelberg (2015)
10. Hürst, W., Snoek, C.G.M., Spoel, W.-J., Tomin, M.: Size matters! how thumbnail number, size, and motion influence mobile video retrieval. In: Lee, K.-T., Tsai, W.-H., Liao, H.-Y.M., Chen, T., Hsieh, J.-W., Tseng, C.-C. (eds.) MMM 2011 Part II. LNCS, vol. 6524, pp. 230–240. Springer, Heidelberg (2011)
11. Hürst, W., Snoek, C.G., Spoel, W.-J., Tomin, M.: Keep moving!: revisiting thumbnails for mobile video retrieval. In: Proceedings of the International Conference on Multimedia, MM 2010, pp. 963–966. ACM, New York (2010)
12. Hürst, W., van de Werken, R.: Human-based video browsing - invesitgating interface design for fast video browsing. In: IEEE ISM 2015 (2015, to appear)
13. Hürst, W., van de Werken, R., Hoet, M.: A storyboard-based interface for mobile video browsing. In: He, X., Luo, S., Tao, D., Xu, C., Yang, J., Hasan, M.A. (eds.) MMM 2015, Part II. LNCS, vol. 8936, pp. 261–265. Springer, Heidelberg (2015)
14. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, MM 2014, pp. 675–678. ACM, New York (2014)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Pereira, F., Burges, C., Bottou, L., Weinberger, K. (eds.) Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates Inc. (2012)
16. Schoeffmann, K.: A user-centric media retrieval competition: the video browser showdown 2012–2014. IEEE MultiMedia **21**(4), 8–13 (2014)
17. Schoeffmann, K., Ahlström, D., Bailer, W., Cobarzan, C., Hopfgartner, F., McGuinness, K., Gurrin, C., Frisson, C., Le, D.-D., Fabro, M., Bai, H., Weiss, W.: The video browser showdown: a live evaluation of interactive video search tools. Int. J. Multimedia Inf. Retrieval **3**, 113–127 (2014)
18. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR, abs/1409.1556 (2014)
19. Uysal, M.S., Beecks, C., Seidl, T.: On efficient content-based near-duplicate video detection. In: CBMI, pp. 1–6 (2015)