WHEN CONTENT-BASED VIDEO RETRIEVAL AND HUMAN COMPUTATION UNITE: TOWARDS EFFECTIVE COLLABORATIVE VIDEO SEARCH

B. Münzer ⋆, *M.J. Primus* ⋆, *M. Hudelist* ⋆, *C. Beecks* •, *W. Hürst* ◦, *K. Schoeffmann* ⋆

⋆ Institute of Information Technology/Lakeside Labs, Klagenfurt University, Austria
{bernd, mprimus, marco, ks}@itec.aau.at
•RWTH Aachen University, Germany (beecks@informatik.rwth-aachen.de)
◦Utrecht University, Netherlands (huerst@uu.nl)

ABSTRACT

Although content-based retrieval methods achieved very good results for large-scale video collections in recent years, they still suffer from various deficiencies. On the other hand, plain human perception is a very powerful ability that still outperforms automatic methods in appropriate settings, but is very limited when it comes to large-scale data collections. In this paper, we propose to take the best from both worlds by combining an advanced content-based retrieval system featuring various query modalities with a straightforward mobile tool that is optimized for fast human perception in a sequential manner. In this collaborative system with multiple users, both subsystems benefit from each other: The results of issued queries are used to re-rank the video list on the tablet tool, which in turn notifies the retrieval tool about parts of the dataset that have already been inspected in detail and can be omitted in subsequent queries. The preliminary experiments show promising results in terms of search performance.

Index Terms— Video retrieval, collaborative search, human computer interaction

1. INTRODUCTION

Content-based video retrieval systems have established as powerful tools for finding specific content in ever-growing large-scale video collections. The increasing interest in such systems is also reflected by public competitions like TRECVid [1] or the annual Video Browser Showdown [2]. A recent survey of this field is provided in [3].

Video retrieval tools are typically built around a retrieval engine that returns a ranked result list according to various multi-modal query features (e.g., by text, example image/clip, sketch, semantic concept, or a combination). In the optimal case the matching video segment(s) should appear at the top of the ranked results. However, video content analysis still suffers from several well-known shortcomings that seriously limit the achievable performance, be it the *semantic gap* [4, 5] or the *usability gap* [6], which refers to the difficulties of users to translate their information need into an appropriate query. Although re-ranking methods and relevance feedback approaches can help to mitigate these problems, it is often the case that the sought video segment(s) still do not appear at the top of the result list. For example, in the *Known-Item Search* (KIS) task of TRECVid [1] in 2010-2012, the wanted video segment did not appear in the Top 100 result list of any of the participating teams for about 29% of all tasks (averaged over all three years, about 9000 videos in total). Another example is the Ad-Hoc search task of TRECVid 2016, where the Median Mean Inferred Average Precision was only 2.4%.

Apart from that, research in the field of video interaction [7] shows that another crucial factor for successful video search is an appropriate interface design. It further shows that even pure human-computation-based approaches can achieve a remarkably high performance at video search. For example, in the Video Browser Showdown (VBS) competition [8] of 2015, a simple but well-designed tablet interface [9] with many small thumbnails of uniformly sampled frames was able to significantly outperform an experienced video retrieval team with a system using sophisticated content search features. The authors argue that providing users with an interface that allows for quick inspection (i.e., human visual filtering) allows to efficiently filter for relevant scenes, even if the data set contains dozens of hours [10]. A similar idea is pursued in the *extreme video retrieval* approach by Hauptmann et al. [11]. However, the practical applicability of such a system is restricted to rather small data sets (up to 25 hours of video content), otherwise the content cannot be inspected in reasonable time. This means that in case of real-world datasets (typically hundreds or thousands of hours), this approach needs some kind of filtering. Such filtering can be provided by content-based retrieval methods. In return, the human-computation approach can help to bridge the semantic gap. To put it in a nutshell, content-based retrieval and human computation can be regarded as complementary approaches that can benefit from each other.

Considering all these aspects, we believe that the key to truly effective video retrieval is to take the best from both worlds and combine the two approaches to a collaborative

978-1-5090-6067-2/17/$31.00

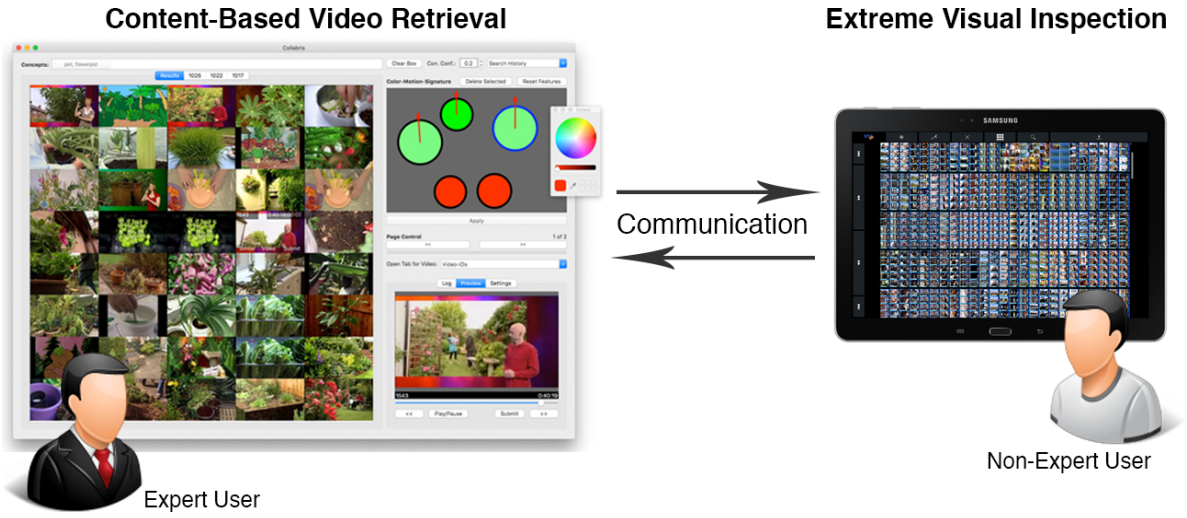


Fig. 1: **Left**: CBVR tool with controls for sketching temporal feature signatures (top right), concept filtering (top left) and tabs for chronological segment browsing. **Right**: Tablet tool, which visualizes video content via uniform sampling of keyframes.

system. The idea of collaborative search is not completely new and has been addressed, e.g., in the work of Smeaton et al. [12]. Their collaborative tabletop system allows pair users to jointly search in the same video data set, while talking, suggesting, discussing, and interacting collaboratively. Their evaluation shows that collaborative search was preferred by users and allowed them to search more effectively together. Nevertheless, research in this field is rather sparse up to now and offers large potential for further research.

In this paper, we present a collaborative video retrieval system that consists of (1) a desktop tool using common video retrieval methods, and (2) a tablet tool focusing on linear search and extreme visual inspection. Moreover, we propose and evaluate two approaches on how the two systems exchange information about retrieved results. The scenario for collaborative usage of such a search system would be a situation with an urgent need for content-based search, such as video search competitions (e.g., VBS), or disaster situations where there is a need for fast search and joint collaboration of several users may help to fulfill this goal faster.

2. COLLABORATIVE VIDEO RETRIEVAL SYSTEM

In this section, we present the two sub-systems of our collaborative video retrieval system and describe how their bidirectional communication is accomplished. Figure 1 shows screen shots of the tools and illustrates their interaction. The basic idea for the collaboration between both tools is to (1) ignore segments in the CBVR tool that have already been inspected by the tablet tool, and (2) to up-rank videos in the tablet tools that have been found in several queries by the CBVR tool. In terms of re-ranking we implemented two different schemes, which are described in detail below.

2.1. CBVR Tool

The CBVR tool uses several common content-based features. Most of the processing is done in advance in order to enable a smooth and responsive user interaction. In the following, we describe the major analysis methods and the user interface.

2.1.1. Shot segmentation

All content-based features that are used in our system are based on shots (also referred to as segments). They are determined by a custom shot detection method based on optical flow tracking. It starts with an initial set of densely sampled points in the frame and uses the Kanade-Lucas-Tomasi (KLT) algorithm [13] to track them from one frame to the next. As soon as the number of trackable points falls below a specific threshold t_C, a shot change is detected and the tracking is restarted with a fresh set of densely sampled points. For each detected shot the middle frame is selected as keyframe.

2.1.2. Concept Detection with CNNs

The keyframes are classified into visual classes that were trained on ImageNet [14]. For that purpose, we employ deep learning with convolutional neural networks (CNN), using the Caffe framework [15]. In particular, we use the "*BVLC AlexNet*" model trained on ILSVRC 2012 data [16], which is freely available on the website of Caffe [15]. We use the five concepts with highest confidence as a result and assign them – together with their confidence – to the corresponding shot, so we can use it for the *query-by-concept* feature. During our experiments we found out that the trained concepts in ILSVRC are not optimally suited for the VBS data and the number of covered concepts varies for each shot, so for future work there is potential for improvement of this component.

2.1.3. Temporal Feature Signatures

Signature-based similarity models have been utilized in many different domains ranging from multimedia data [17, 18] to scientific data [19, 20]. A similar system in terms of such models has been shown by Blaŝek et al. [21, 22]. With their system they won the VBS competition in 2014 and 2015 [8]. In our system, we use an extension of the feature signature model proposed in [23] that additionally take into account temporal characteristics of features. In particular, they facilitate dynamic shot-wise content aggregation by utilizing object-specific feature quantizations. We use these temporal feature signatures for similarity search as well as for the "query-by-sketch" modality.

We model the content-based properties of each single keyframe 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 [24], 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 keyframe, 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 details see [23]).

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 keyframe 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 Figure 2).

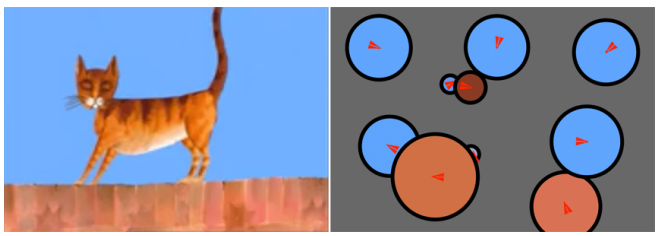


Fig. 2: Right: example visualization of a temporal feature signature (TFS) for a given keyframe (left). It is important to note that the TFS aggregates information from the entire shot, not only the keyframe.

Based on the temporal feature signatures described above, an asymmetric variant of the *Signature Matching Distance* [25] 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.

2.1.4. User Interface

The user interface of the CBVR tool provides various options to filter the set of video segments. Retrieval results are always displayed in a new tab, so that several search strategies can be pursued at the same time (similar to the *facet-based search* approach proposed in [26]). Additionally, a search history allows to revisit any previous search result. Each video segment is represented by the middle frame, but more detailed information about the shot can be quickly inspected by hovering over the image. Moreover, each segment can be played back in a dedicated preview section.

- **Query-by-concept:** Users can filter the data set for matching concepts by a concept name. It is also possible to define the minimum confidence for detected semantic concepts to be included.

- **Query-by-sketch:** Users can define spatial or temporal *feature signatures* by drawing a sketch. Clicking anywhere in the box creates a new cluster, which can be adjusted in terms of color, location, and size. A right mouse click allows to draw a motion vector for a cluster, to be used for matching with the temporal feature signature (if no motion vector is specified, only spatial matching is performed). These color sketches can be combined with selected concepts too.

- **Query-by-example:** The user interface provides a search-by-example feature, where the most similar shots for a selected segment can be retrieved, based on the *Signature Matching Distance* [25] of underlying temporal feature signatures.

- **Browsing:** For situations where users would like to start search by browsing instead of querying, the interface allows to open chronological lists of shots for selected videos.

2.2. Tablet Tool

The user interface of the tablet tool uses a simple storyboard layout that is optimized for linear inspection of a large number of very small thumbnails, as proposed in [27]. More specifically, it visualizes the whole data set by temporally arranged thumbnails, in rows of thumbnails presented in column-major order for better visual coherence. It does not rely on video shots, but uses videos as basic entities. The thumbnails are uniformly sampled from the videos at a rate of 1 fps. In this way, the tablet tool displays up to 625 images at a single page (see Figure 1) and provides buttons to jump one page up or down. The number of thumbnails on a single page has been chosen according to the findings of [28] and is a tradeoff between providing maximum amount of overview while keeping the thumbnails large enough to recognize visual details.

2.3. Collaboration

As already mentioned, the main idea of our collaborative system is to combine the strengths of both tools while mitigating their weak spots. On the one hand, the CBVR tool is very good at performing a coarse filtering of the potentially huge data collection by ranking video shots according to their relevance based on various query modalities. However, this ranking usually is not precise enough for finding the wanted shot(s) at the very first positions of the result list. Therefore, a user would often rather formulate a new query instead of inspecting several pages of results, although the correct segment might be contained therein. If we assume that a user always looks at the first 100 results (of 10000's), it might happen that the wanted shot is ranked on position 101. This might even happen for several different queries. In this example, the performance of the system is actually very good, but the user still would not find the desired shot. Hence, it would be desirable to aggregate the result lists of different queries to one list that should be inspected in detail. On the other hand, the tablet tool is able to exploit superior human perception capabilities for detailed inspection, but only for a restricted amount of data. In preliminary tests we discovered that in five minutes a typical user is able to carefully inspect about 25 hours of video content. For larger video collections, it is essential to have a mechanism that re-ranks the video list in a way that the relevant video(s) are included in these first 25 hours. Consequently, our approach for collaboration between these two complementary tools is as follows:

- Shots from videos that already have been inspected on the tablet tool are down-ranked or omitted in query results of the CBVR tool (for the current search session).
- The order of the videos visualized on the tablet tool is continuously updated based on the results of queries issued at the CBVR tool. Hence, videos that frequently occur on top positions are up-ranked and have a higher chance of being inspected by the tablet user.

While omitting shots on the CBVR tool is straightforward, the conception of an appropriate re-ranking strategy poses an interesting research question. In the following, we present two different approaches for this problem.

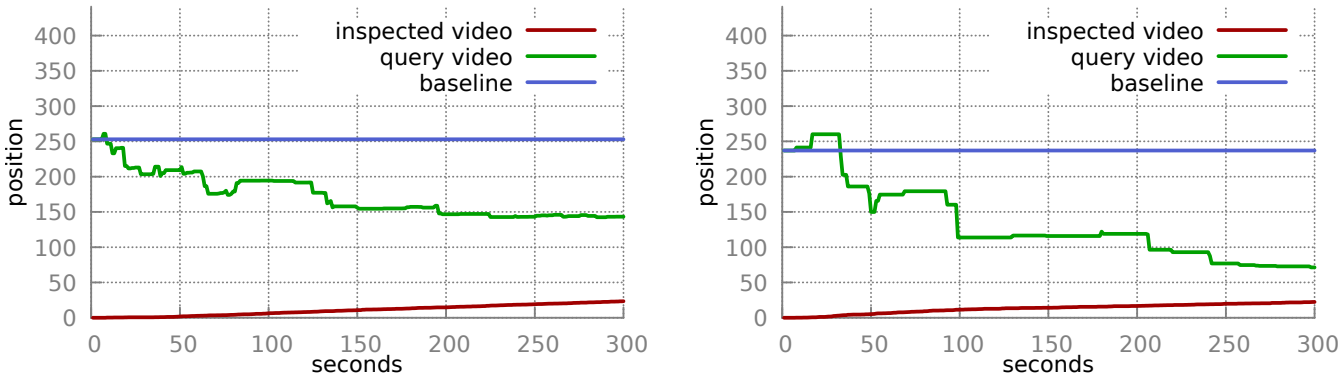
1. The **automatic approach** focuses on an immediate and continuous exchange of information between the two systems, without any intervention of the user. The tools exchange information regarding viewed videos and query results in the background. Every time the desktop user applies a query, the top 250 results of this query are sent to the tablet tool and used to re-rank the remaining list of videos. This list of query results contains a ranked list of shot IDs. The tablet tool counts how many matching shots belong to a particular video and use this *result count* of every video to rank the non-inspected videos in descending order.

2. A problem of the automatic approach might be that usually not every query by the expert user returns meaningful results. In our experiments we noticed that users of the CBVR tool often experiment with the filtering options until they find a promising setting. Therefore, we also propose a **manual approach** that requires explicit action of the expert user to send results of a query to the tablet for re-ranking. Our hypothesis is that this manual approach will provide a better ranking on the tablet, since we avoid noise from unsuccessful queries.

3. EXPERIMENTAL RESULTS

In order to evaluate our collaborative video retrieval approach we performed some user experiments. The setup is based on the challenge posed in the Video Browser Showdown competition [8]: a target video segment with a duration of 20 seconds is presented on a large screen. The participants have to find this segment in a data set of hundreds of hours of video content. This modality is generally known as "Known-item search". We also used the data set of the 2016 VBS competition. It contains 441 video files with a total duration of about 250 hours. The test was performed in our lab with 6 expert and novice users that collaboratively performed 30 search tasks. Similarly to the VBS we defined a maximum search time of five minutes per task.

Our experiments show that both re-ranking approaches support the tablet user very well, as they considerably increase the chance to reach the target video (as illustrated in Figure 3). However, we observed that the automatic approach frequently causes the target video to unpredictably fluctuate in the video list, without reaching the realistically inspectable range of videos on the tablet tool (which is about 25 hours of content). This can be explained by the way desktop users experiment with various filtering options of the CBVR tool. Since the first set of queries barely create an optimal result



(a) *Automatic* re-ranking approach.

(b) *Manual* re-ranking approach.

Fig. 3: Averaged results of the two proposed re-ranking approaches.

set users tend to optimize their settings during the search process in an iterative way. Sometimes, they even start over and try a completely different approach. As the tablet tool is influenced by each and every query the CBVR tool processes, the re-ranking gets ”polluted”. This problem is illustrated in Figure 3(a), where the green line represents the position of the target video in the video list and the red line represents the current position of the tablet user (averaged over all users and tasks). The green and red line converge very slowly and do not intersect by the end of the task. Nevertheless, the automatic approach considerably improves the rank of the target video compared to the ”baseline” approach without any re-ranking (from 250 of 441 total videos to 143, which corresponds to an improvement of 43%).

The manual approach performs even better. It not only changes the way the systems communicate, it also changes the way users cooperate. As desktop users now have the control over what is sent to the tablet tool, they are much more sensitive about their filtering settings. As can be seen in Figure 3, the target video is re-ranked to position 113 already after 100 seconds on average, while this position is not even reached after five minutes with the automatic approach. The median final position of the manual re-ranking approach is 71. This corresponds to an impressive improvement of 70%, while the automatic approach only reaches 43%.

4. CONCLUSION

In this paper we present a new concept of collaborative video search, which combines the advantages of content-based retrieval and human computation through information exchange about the search status. We propose two approaches for re-ranking the results on a tablet tool and show that both approaches help to achieve a much better performance than without re-ranking. The performed user experiments show that the manual approach clearly outperforms the automatic approach. It better supports the mobile user because it is more reliable in providing a list of relevant videos and thus significantly increases the chance to hit the target video. We found out that the reason for this is the fact that automatic communication introduces too much noise. Thus, we conclude that it is more effective to perform a re-ranking based only on explicit input of the expert user who operates the CBVR tool.

For now, we only considered a known-item search scenario, but in future work we intend to apply our approach also for ad-hoc search, which allows multiple correct answers. We expect that our approach performs even better in this setting. A further potential for future work is to incorporate multiple tablet users in the collaboration team and distribute the re-ranked video list over multiple instances of the tablet tool to further improve the overall retrieval performance.

5. 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.

6. REFERENCES

[1] A. F. Smeaton, P. Over, and W. Kraaij, “Evaluation campaigns and trecvid,” in *MIR ’06: Proc. of the 8th ACM Int’l Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[2] K. Schoeffmann, D. Ahlström, W. Bailer, C. Cobârzan, F. Hopfgartner, K. McGuinness, C. Gurrin, C. Frisson, D.-D. Le, M. Del Fabro, H. Bai, and W. Weiss, “The Video Browser Showdown: a live evaluation of interactive video search tools,” *Int. Journal of Multimedia Information Retrieval*, pp. 1–15, 2013.

[3] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 6, pp. 797–819, 2011.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comp. Surv.*, vol. 40, no. 2, pp. 5:1–5:60, 2008.

[5] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1349–1380, Dec 2000.

[6] K. Schoeffmann and F. Hopfgartner, "Interactive video search," in *Proc. 23rd ACM Int'l Conf. on Multimedia*, New York, USA, 2015, MM '15, pp. 1321–1322, ACM.

[7] K. Schoeffmann, M. A. Hudelist, and J. Huber, "Video interaction tools: A survey of recent work," *ACM Comput. Surv.*, vol. 48, no. 1, pp. 14:1–14:34, Sept. 2015.

[8] K. Schoeffmann, "A user-centric media retrieval competition: The video browser showdown 2012-2014," *MultiMedia, IEEE*, vol. 21, no. 4, pp. 8–13, Oct 2014.

[9] W. Hürst and M. Hoet, "Sliders versus storyboards - investigating interaction design for mobile video browsing," in *MultiMedia Modeling*, vol. 8936 of *LNCS*, pp. 123–134. Springer, 2015.

[10] W. Hürst and R. v. d. Werken, "Human-based video browsing - investigating interface design for fast video browsing," in *2015 IEEE ISM*, 2015, pp. 363–368.

[11] A. G. Hauptmann, W.-H. Lin, R. Yan, J. Yang, and M.-Y. Chen, "Extreme video retrieval: joint maximization of human and computer performance," in *Proc. 14th ACM Int'l Conf. on Multimedia*. ACM, 2006, pp. 385–394.

[12] A. F. Smeaton, H. Lee, C. Foley, and S. McGivney, "Collaborative video searching on a tabletop," *Multimedia Systems*, vol. 12, no. 4-5, pp. 375–391, 2007.

[13] J.-Y. Bouguet, "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm," *Intel Corporation*, vol. 5, no. 1-10, pp. 4, 2001.

[14] 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*, June 2009, pp. 248–255.

[15] 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*, New York, NY, USA, 2014, MM '14, pp. 675–678, ACM.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Sys. 25*, pp. 1097–1105. Curran Associates, Inc., 2012.

[17] C. Beecks, S. Kirchhoff, and T. Seidl, "On stability of signature-based similarity measures for content-based image retrieval," *MTAP*, vol. 71, no. 1, pp. 349–362, 2014.

[18] M. S. Uysal, C. Beecks, and T. Seidl, "On efficient content-based near-duplicate video detection," in *CBMI*, 2015, pp. 1–6.

[19] M. S. Uysal, C. Beecks, J. Schmücking, and T. Seidl, "Efficient similarity search in scientific databases with feature signatures," in *SSDBM*, 2015, pp. 30:1–30:12.

[20] C. Beecks, M. Hassani, J. Hinnell, D. Schüller, B. Brenger, I. Mittelberg, and T. Seidl, "Spatiotemporal similarity search in 3d motion capture gesture streams," in *SSTD*, 2015, pp. 355–372.

[21] J. Lokoĉ, A. Blaŝek, and T. Skopal, "Signature-based video browser," in *MultiMedia Modeling*, vol. 8326 of *LNCS*, pp. 415–418. Springer, 2014.

[22] A. Blaŝek, J. Lokoĉ, F. Matzner, and T. Skopal, "Enhanced signature-based video browser," in *MultiMedia Modeling*, vol. 8936 of *LNCS*, pp. 243–248. Springer, 2015.

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

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

[25] C. Beecks, S. Kirchhoff, and T. Seidl, "Signature matching distance for content-based image retrieval," in *Proc. of the 3rd ACM int'l conference on multimedia retrieval*. ACM, 2013, pp. 41–48.

[26] R. Villa, N. Gildea, and J. M. Jose, "Facetbrowser: a user interface for complex search tasks," in *Proc. 16th ACM Int'l Conf. on Multimedia*, 2008, pp. 489–498.

[27] W. Hürst, R. van de Werken, and M. Hoet, "A storyboard-based interface for mobile video browsing," in *MultiMedia Modeling*, vol. 8936 of *Lecture Notes in Computer Science*, pp. 261–265. Springer, 2015.

[28] W. Hürst, C. G. M. Snoek, W. J. Spoel, and M. Tomin, "Size matters! how thumbnail number, size, and motion influence mobile video retrieval," in *International Conference on MultiMedia Modeling*, 2011, pp. 230–240.