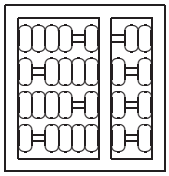


INSTITUTO DE COMPUTAÇÃO
UNIVERSIDADE ESTADUAL DE CAMPINAS

Intuitive Video Browsing Along Hierarchical Trees

Jurandy Almeida *Sheila M. Pinto Cáceres*
Ricardo da S. Torres *Neucimar J. Leite*

Technical Report - IC-11-06 - Relatório Técnico

March - 2011 - Março

Intuitive Video Browsing Along Hierarchical Trees

Jurandy Almeida Sheila M. Pinto Cáceres Ricardo da S. Torres
Neucimar J. Leite

Institute of Computing, University of Campinas – UNICAMP
13083-852, Campinas, SP – Brazil
{jurandy.almeida,rtorres,neucimar}@ic.unicamp.br,
sheila.caceres@students.ic.unicamp.br

March 3, 2011

Abstract

Recent advances in technology have increased the availability of video data, creating a strong requirement for efficient systems to manage those materials. Making efficient use of video information requires that data be accessed in a user-friendly way. Ideally, one would like to perform video search using an intuitive tool. Most of existing browsers for the interactive search of video sequences, however, have employed a too rigid layout to arrange the results, restricting users to explore the results using list- or grid-based layouts. In this paper, we present a novel approach for the interactive search that displays the result set in a flexible manner. The proposed method is based on a hierarchical structure called *Divisive-Agglomerative Hierarchical Clustering* (DAHC). It is able to group together videos with similar content and to organize the result set in a well-defined tree. This strategy makes the navigation more coherent and engaging to users.

1 Introduction

Advances in data compression, data storage, and data transmission have facilitated the way videos are created, stored, and distributed. The increase in the amount of video data has enabled the creation of large digital video libraries. This has spurred great interest for systems that are able to efficiently manage video material [8, 14, 25].

Making efficient use of video information requires that data to be accessed in a user-friendly way. For this, it is important to provide users with a browsing tool to interactively search for (or query) a video in large collections, without having to look through many possible results at the same time, so that a user can easily find the video in which he/she is interested.

A lot of research has being done in browsing techniques for the interactive search of video sequences [10–13, 26–28]. However, many of those research works have considered a rigid layout to arrange the result set in some default order, typically according to the relevance to the query.

In this paper, we present a novel approach for the interactive search that displays the result set in a more flexible and intuitive way. It relies on a hierarchical structure, where visually or semantically alike videos are placed together. Such a structure is achieved by a clustering algorithm known as *Divisive-Agglomerative Hierarchical Clustering* (DAHC) [5, 7,22], which is able to organize the results in a well-defined tree. This innovative framework is significant different from traditional paradigms, which often limit users to explore the results using list- or grid-based layouts.

The remainder of this paper is organized as follows. Section 2 introduces the background of interactive search problems. Section 3 describes related work. Section 4 presents our approach and shows how to apply it for browsing a large video collection. Finally, we offer our conclusions and directions for future work in Section 5.

2 Background

The exploration of large collections of video data is non-trivial. When a user requests a search, the query formulation (search criterion) can be quite difficult.

Most of search systems are based on textual metadata, which leads to several problems when searching for visual content. Generally, the user lacks information about which keywords best represent the content that he/she is interested. In fact, different users tend to use different words to describe a same visual content. The lack of systematization in choosing query words can significantly affect the search results [12].

Modern systems have addressed those shortcomings by automatically detecting visual concepts derived from visual properties, such as color, texture, and shape. However, a minimum knowledge about the concept vocabulary is needed for performing a query, which is not appropriate for non-expert users [27].

Fully automated approaches have combined descriptors of multiple modalities (textual metadata, visual properties, and visual concepts). In spite of all the advances, the formulation of a query using such features is a difficult task for a human interested in a specific video [13].

Once the search results are returned, we can explore many different directions based on query type and user intention. Several visualization techniques have been proposed to assist users in the exploration of result sets [11, 17, 21, 26].

Those methods often employ dimensionality reduction algorithms to map the high-dimensional feature space of visual properties into a fixed display. Afterwards, a display strategy is applied for producing user-browsable content [28].

There are two basic kinds of navigation [12]: targeted search and exploratory search. The former performs a fast browsing in a single list of results. The latter allows the user to control the browsing procedure in several ways.

The major challenge of designing an interactive display is the fatigue and frustration that a user might experience. In general, users can spend a limited time to identify relevant videos for a query, thus they are hard-pressed to quickly inspect a large set of results.

The layout of videos is another concern for an interactive system. An effective tool for browsing in large collections should be suited for users without any expertise, providing an easy way to use the interface.

3 Related Work

In this work, we are interested in visualization techniques for the interactive search. A comprehensive review of browsing models can be found in [16]. Some of the main ideas and results among the previously published methods are briefly discussed next.

Adcock et al. [1] organize videos from the result set into a storyboard-based interface. In this way, a user can quickly evaluate the relevance of a whole section of results.

Hauptmann et al. [15] arrange the result set in a grid. They present pages dynamically to users in a very rapid fashion, allowing them to effectively explore the results at high speed.

Zavesky et al. [27,28] introduced the notion of visual islands. They also use a grid-based display to dynamically organize groups of similar keyframes in every page of results.

Rautiainen et al. [21] combine both the timeline of videos and the visual similarity in a single view. They dispose the search results in a grid which vertically align similar shots for a given keyframe of the timeline.

Snoek et al. [23] also show the video timeline in the horizontal dimension. However, instead of using a full grid, the results for the current shot of the timeline are ranked in a vertical list.

De Rooij et al. [11,12] extended the Snoek's approach by including up to eight different ranked lists. In addition, they display the visual and conceptual similarities of video shots in separate dimensions, allowing users rather explore the results.

Heesch et al. [17] organize both the time and the similarity of videos in a directed graph. Thus, a user can refine the result set by controlling the weight of each arc, which enables many navigation possibilities.

In general, all the previous works employ a list- or grid-based layout for browsing the result set. Different from all those techniques, our approach organize the results in a hierarchical structure which group together videos with a similar content. This strategy allows users to explore the result set in a more flexible and intuitive way.

4 Our Approach

Save time in browsing, intuitively comprehend the results, and allow an exploratory search: those are the basic principles of our approach. In the following subsections, different design choices to achieve such goals are discussed in more details.

4.1 Features and Similarity

Humans judge more quickly the relevance of interrelated items. However, discovering the ideal relationship for such a judgement is non-trivial. The simplest approach is to group together video frames with similar content, so that a relevant judgement for one video frame could be applied to all near-duplicated ones and, hence, maximize the diversity.

In our approach, stories are the meaningful and manageable units for presenting the result set to the user. They consist of multiple shots and are represented by a collection of frames, as illustrated in Figure 1. This strategy provides an easy way for the user to visually judge whether a story is worth exploring.

Figure 1: An example of storyboard produced for the video *Senses And Sensitivity, Introduction to Lecture 3 presenter*.

We use the method described in [6] to build storyboards. A flowchart of this approach is shown in Figure 2. For each frame of an input sequence, visual features are extracted from the video stream for describing its visual content. After that, a simple and fast algorithm is used to detect groups of video frames with a similar content and for selecting a representative frame per each group. Finally, the selected frames are filtered in order to avoid possible redundant or meaningless frames in the storyboard.

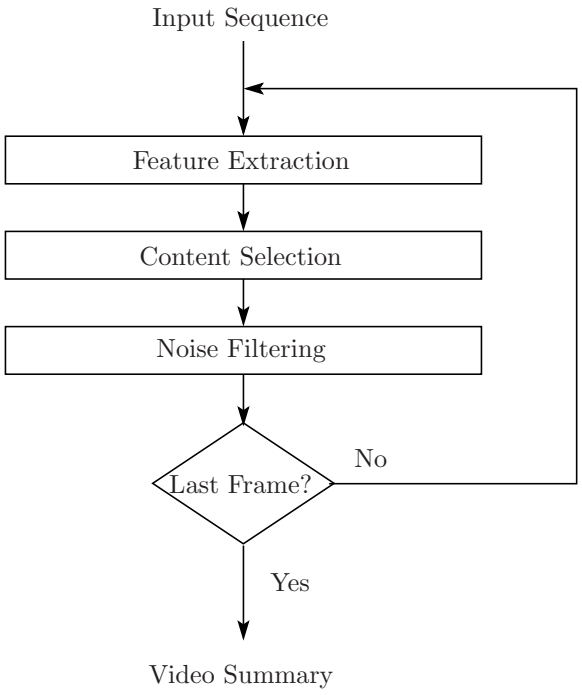


Figure 2: Flowchart of the method used to build storyboards [6].

Numerous forms of raw features can be used to determine the similarity between video frames of different storyboards. Each type of feature spans a multidimensional feature space. The distance function determines the dissimilarity between features within this space. Thus, we coordinate the display positions of each story using its dissimilarity space.

Our technique was designed to be flexible and robust and, therefore, the feature input is not limited to any one type. Instead, all possible data types can be used. The only requirement is that the dissimilarity between features must be numerically represented by an appropriate distance metric.

4.2 Intuitive Display

A problem regarding the interactive search of video sequences is the human understanding of what the system was trying to judge as relevant.

The most common approach for designing an interactive display is to use dimensionality reduction algorithms to map the multidimensional feature space into a fixed display and to apply a display strategy for producing user-browsable content.

Advanced visualization techniques use clustering algorithms to divide the result set into semantic units and, hence, enhance the speed at which the user is able to analyze the results.

In general, clustering methods can be partitional or hierarchical. Partitional algorithms typically determine all clusters at once. In contrast, hierarchical algorithms find successive clusters using previously established ones. These algorithms usually are either agglomerative (bottom-up) or divisive (top-down). Agglomerative algorithms begin with each element as a separate cluster and merge them into successively larger clusters. Divisive algorithms begin with the whole set and proceed to divide it into successively smaller clusters [18].

Our approach adopt a hybrid algorithm, called *Divisive-Agglomerative Hierarchical Clustering* (DAHC) [5, 7, 22]. It combines the advantages of both divisive and agglomerative paradigms in order to achieve a set of high quality clusters, which are organized in a well-defined tree structure.

Consider an initial set of results. First, it is divided into groups based on their global distribution. This partitioning can be further refined by dividing each existing group based on the local distribution of a subset of results. Such a process may be repeated by taking a smaller subset at each time until no further improvements are possible. Finally, we have a hierarchical set of groups. This is roughly the basic idea of DAHC. In other words, given a query, the search space can be reduced by gradually considering a subset of results with a more relevant distribution.

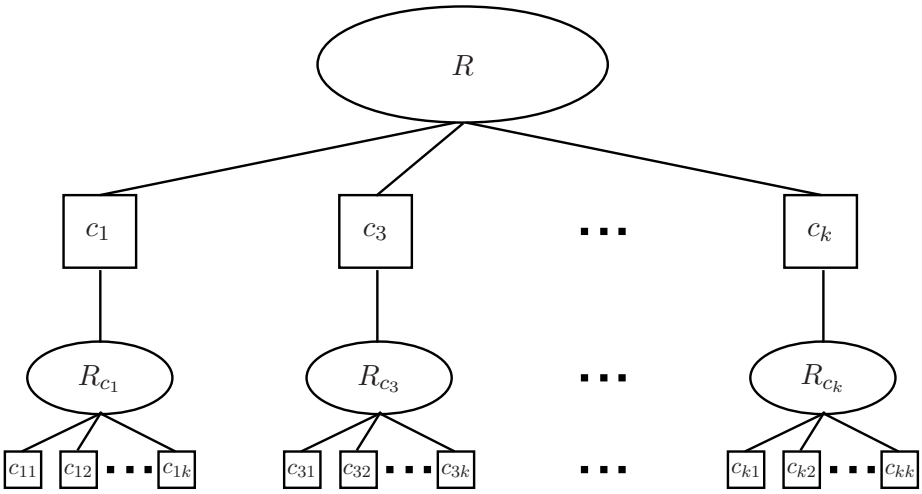


Figure 3: A representation of DAHC.

The tree construction is performed in a top-down fashion. In order to clarify this approach, look at Figure 3. At the beginning, the set of results $R = \{r_1, r_2, \ldots, r_N\}$ is

considered to be part of a single cluster. Results in this set are first divided into $k \leq K$ clusters $c_1, c_2, \ldots, c_k$. For each group c_i, a subset R_{c_i} is created by grouping the results of c_i and those of f adjacent clusters. To build subsequent levels of the tree, this process of dividing and grouping is repeated for all of the new subset of results at each level, creating a hierarchical structure. The process stops when the number of results in a subset is less than or equals to K or the number of clusters spanned by a subset is less than the double of f. For a detailed discussion of this procedure, refer to [22].

The key advantage of our technique is to visualize the result set by displaying the stories based on the fractal shape of a tree, as illustrated in Figure 4. This strategy allows the user to view the relationship between several clusters at once.

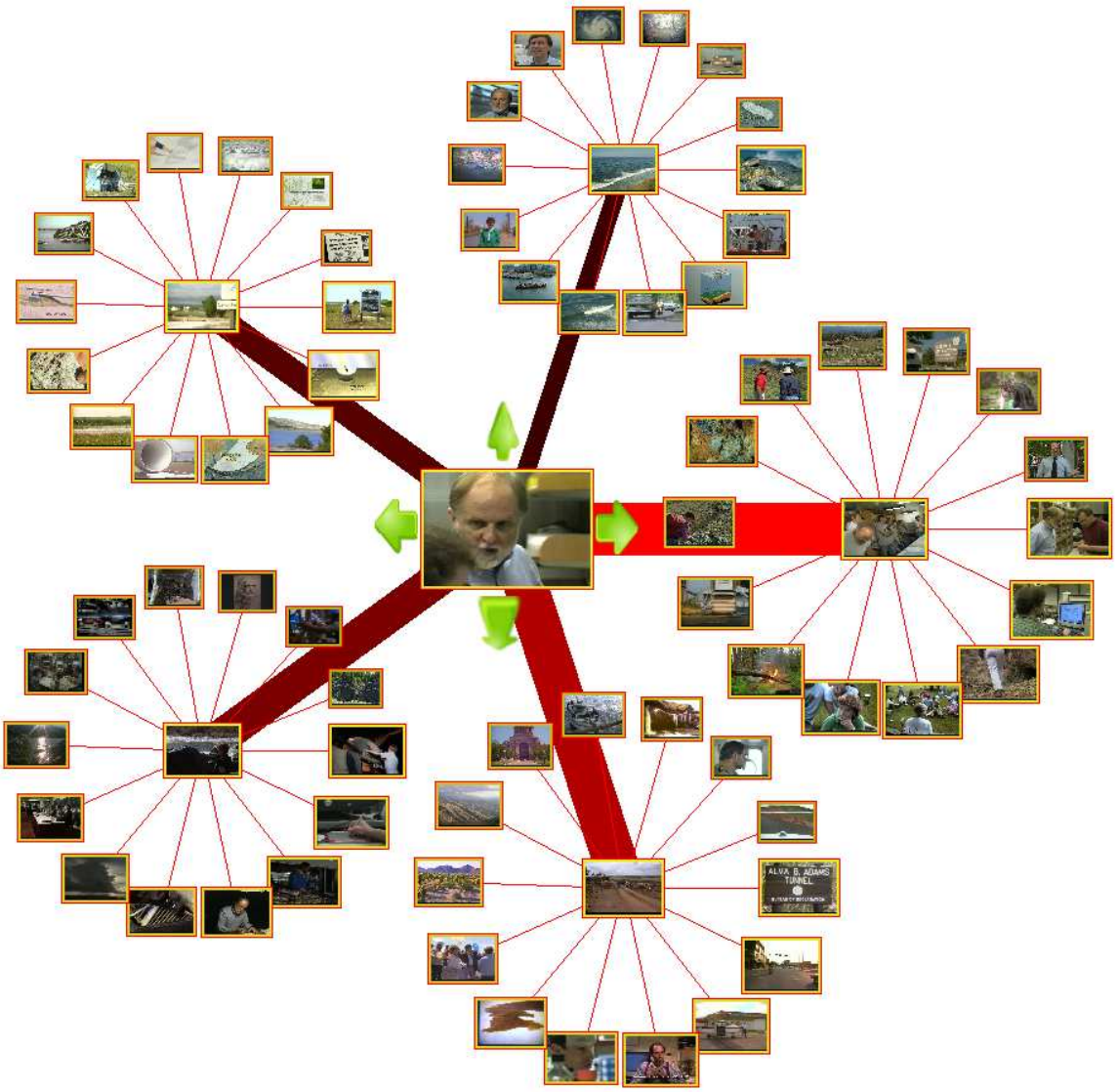
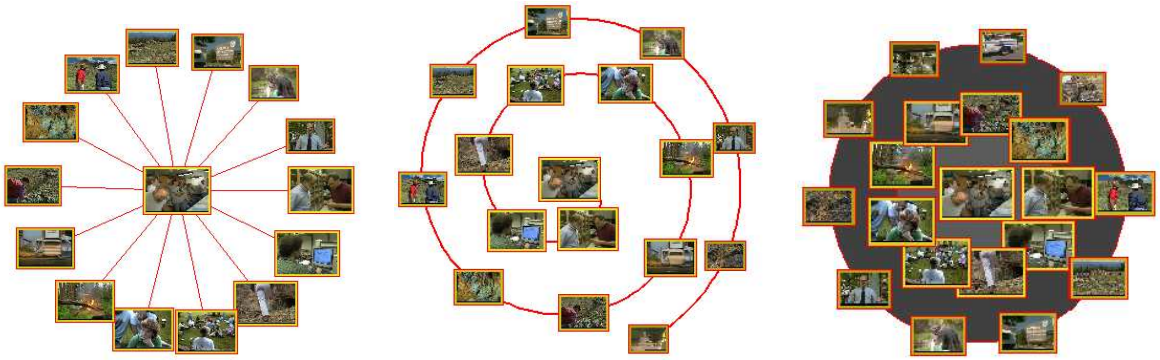


Figure 4: An example of the visualization of our approach.



(a) Circular (b) Spiral (c) Concentric Rings

Figure 5: Three visualization strategies provided by our approach.

In this figure, we display the stories for a sample of 50 videos randomly selected from the Open Video Project[1]. All videos are in MPEG-1 format (at 352 × 240 resolution and 29.97 frames per second), in color and with sound. The selected videos are distributed among several genres (e.g., documentary, educational, ephemeral, historical, lecture) and their duration varies from 1 to 4 minutes. Those videos are the same used in [6] and their storyboards can be seen at http://www.liv.ic.unicamp.br/~jurandy/summaries/.

We converted each frame of those storyboards to a 64-dimensional feature vector by computing a Color Histogram [24]. The color histograms were extracted as follows: the RGB space is divided into 64 subspaces (colors), using four ranges in each color channel. The value for each dimension of a feature vector is the density of each color in the entire frame. The distance function used to compare the feature vectors is the Manhattan (L_1) distance.

Our approach places the query in the center of the visualization display. Thus, we force the user to focus his/her attention on the center of the screen. The clusters of a given level are circularly distributed around the query in a clockwise order of similarity regarding the query, which is represented by the width of the connecting line between them.

This strategy is also applied to each of the clusters. In this way, the user has a more intuitive understanding of the display. At the center, we present the most relevant result. It represents the centroid of its cluster. The remaining results are sorted in a circular manner according to a clockwise order of similarity with respect to the centroid, which is denoted by their size and border color (in a color gamma from yellow to dark green).

In this way, we provide a coherent distribution of the query-related video universe by setting the results over a well-organized structure. This distribution, in most cases, avoid overlapping, which represents a valuable advantage over other cluster-based visualization techniques [9, 19, 20].

Moreover, our technique was designed to provide a totally flexible display. Thus, the user can also iterate with the visualization for producing different views of the result set. It can be applied to all the levels of the visualization, allowing for the creation of hybrid structures. In this way, the proposed method maximize the useful information delivered to users more quickly.

[1]http://www.open-video.org/

Figure 5 presents three possible visualization strategies for each cluster. The first technique (see Figure 5(a)) distributes the results circularly, as explained above. The second method (see Figure 5(b)) organizes the results over a spiral structure. The centroid is located at the origin of the spiral and successive results are distributed over the spiral line in increase order of relevance. In the third scheme (see Figure 5(b)), the results are arranged as a series of concentric rings. Similar results with respect to the centroid are located over a nearer ring.

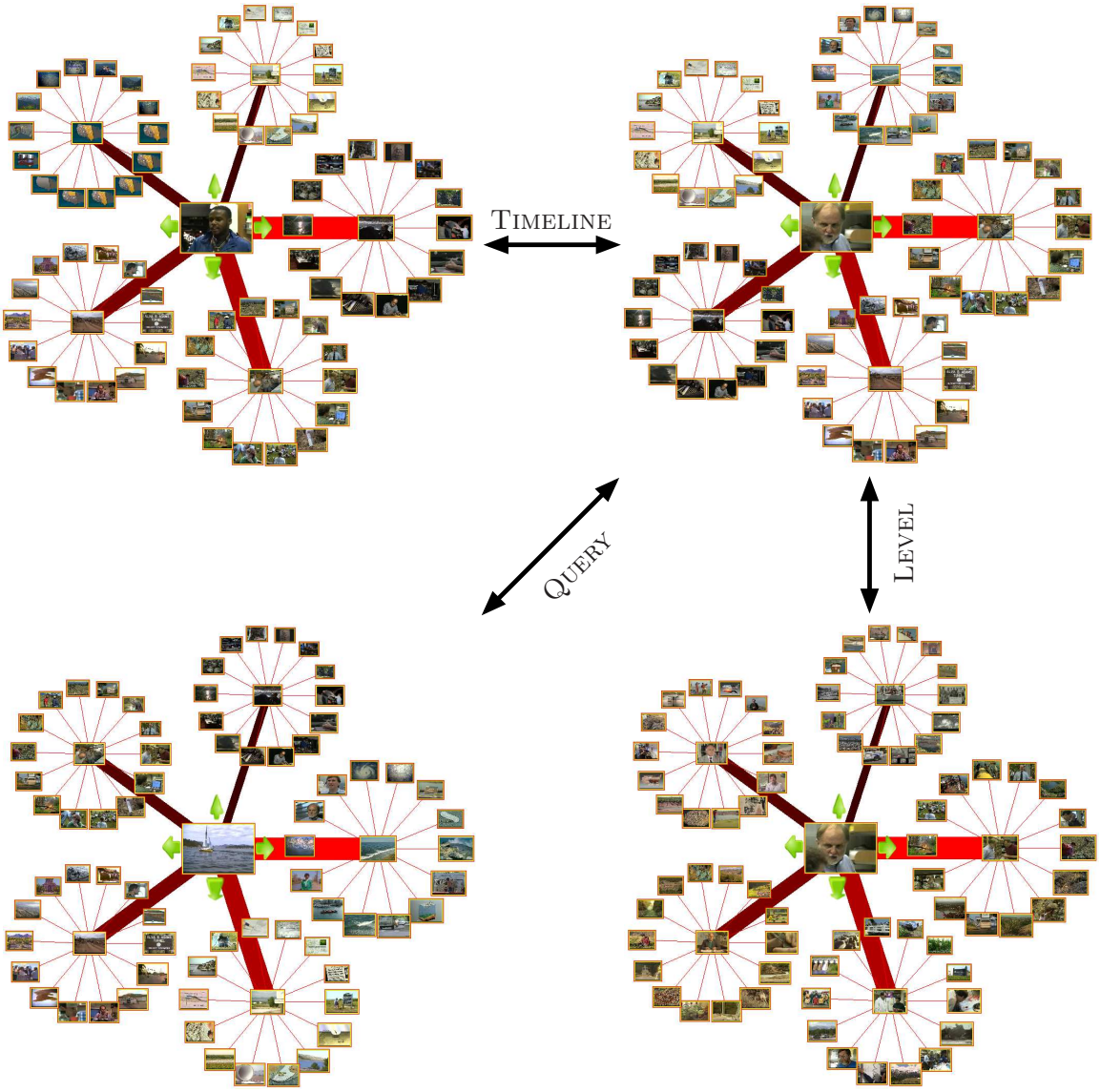


Figure 6: Overview of the navigation options of our approach.

4.3 Engaged and Guided Browsing

The fully utilization of a user's inspection ability requires an engaging display which is guided by user preferences. Our approach fulfill such a principle by dynamically rearranging the result set. Figure 6 presents the navigation options of the propose method. Those options indicate all the possible browsing directions of a user.

Using a mouse click or a key press, the user can give an indication of which story is the most relevant. Then, we place the user at a new set of results which is most related to the last story marked as relevant, as illustrated in the right column of the Figure 6. For that, we follow a path down the tree, descending to the subtree containing that story. By iteratively pruning subtrees, the user can efficiently explore the result set to quickly find relevant results.

Our approach is totally flexible, allowing users to navigate laterally in the timeline of the video. Thus, the result set is updated whenever they decide to focus in another story of a video. The top line of the Figure 6 illustrates such a transition. In this way, the user can combine both targeted search and temporal browsing, which often yields more relevant results.

Anytime users can also change the video-of-interest by choosing any of the stories visible in the screen. Thus, they can interleave between targeted search and exploratory search. This changing is illustrated in the second diagonal of the Figure 6. Using different searching and browsing methods into a single environment enhance the user's inspection ability. In this way, the user is in complete control and can change the current view at any time.

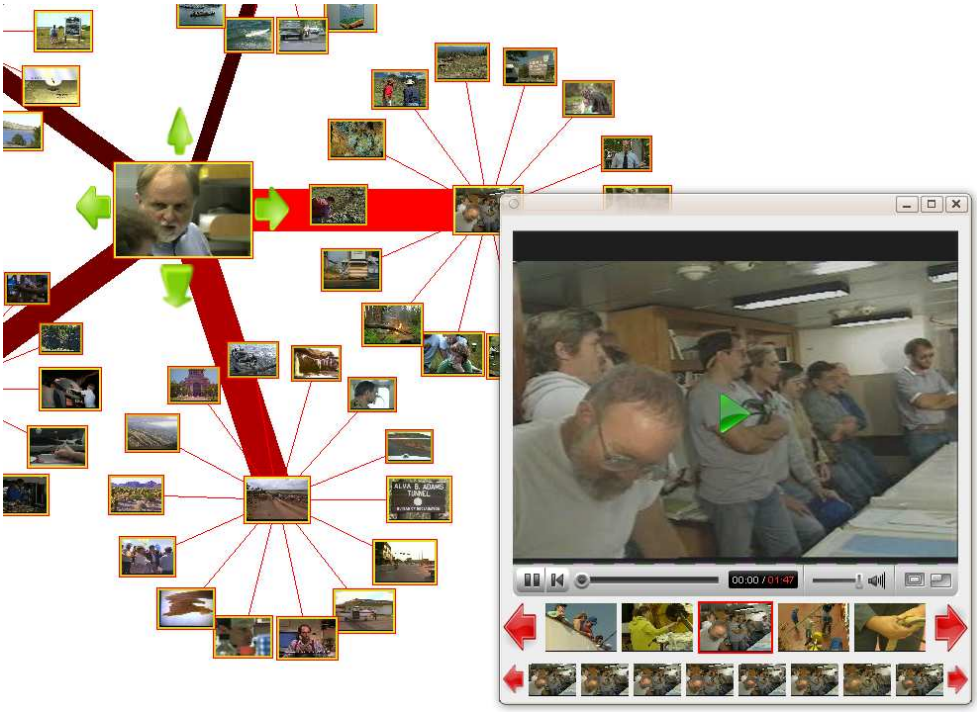


Figure 7: The pop-up window where a specific story is handled.

Additionally, we integrate different functionalities in the interface. By right-clicking on the stories, the user is presented with the operations that can be performed to them. This opens a pop-up window on the screen, as illustrated in Figure 7. On the top, we make available a video player. Below, we display a click-able sequence of story collages and the selected story is highlighted in the video timeline. The frames from the shots in the selected story are expanded on the bottom.

5 Conclusions

In this paper, we have presented a novel approach for the interactive search that displays the result set in a more flexible and intuitive way. Our technique relies on a hierarchical structure, where visually or semantically alike videos are placed together. Such a strategy offers a guided browsing more coherent and engaging to users. These benefits were carefully demonstrated in our showcases.

Future work includes a subjective evaluation of our approach. In addition, we also want to evaluate other visual features and similarity metrics. Moreover, the proposed method can be augmented to consider local features [4] and/or motion analysis [2, 3]. Finally, we plan to investigate the effects of integrating our technique into a complete system for search-and-retrieval of video sequences.

Acknowledgment

This research was supported by Brazilian agencies FAPESP (Grant 07/52015-0, 08/50837-6, 09/04732-0, and 09/18438-7), CNPq (Grant 311309/2006-2, 472402/2007-2, 135526/2008-6, and 306587/2009-2), and CAPES (Grant 01P-05866/2007).

References

[1] J. Adcock, M. L. Cooper, A. Girgensohn, and L. Wilcox. Interactive video search using multilevel indexing. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'05)*, pages 205–214, 2005.

[2] J. Almeida, R. Minetto, T. A. Almeida, R. S. Torres, and N. J. Leite. Robust estimation of camera motion using optical flow models. In *Proceedings of the International Symposium on Advances in Visual Computing (ISVC'09)*, pages 435–446, 2009.

[3] J. Almeida, R. Minetto, T. A. Almeida, R. S. Torres, and N. J. Leite. Estimation of camera parameters in video sequences with a large amount of scene motion. In *Proceedings of the IEEE International Conference on Systems, Signals and Image Processing (IWSSIP'10)*, pages 348–351, 2010.

[4] J. Almeida, A. Rocha, R. S. Torres, and S. Goldenstein. Making colors worth more than a thousand words. In *Proceedings of the ACM International Symposium on Applied Computing (ACM SAC'08)*, pages 1180–1186, 2008.

[5] J. Almeida, R. S. Torres, and N. J. Leite. BP-tree: An efficient index for similarity search in high-dimensional metric spaces. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'10)*, pages 1365–1368, 2010.

[6] J. Almeida, R. S. Torres, and N. J. Leite. Rapid video summarization on compressed video. In *Proceedings of the IEEE International Symposium on Multimedia (ISM'10)*, pages 113–120, 2010.

[7] J. Almeida, E. Valle, R. S. Torres, and N. J. Leite. DAHC-tree: An effective index for approximate search in high-dimensional metric spaces. *Journal of Information and Data Management*, 1(3):375–390, 2010.

[8] S.-F. Chang, W. Chen, H. J. Meng, H. Sundaram, and D. Zhong. A fully automated content-based video search engine supporting spatio-temporal queries. *IEEE Transactions on Circuits Systems and Video Technology*, 8(5):602–615, 1998.

[9] C. Chen, G. Gagaudakis and P. Rosin. Content-based image visualization. In *Proceedings of the International Conference on Information Visualisation (IV'00)*, pages 13–18, 2000.

[10] M. G. Christel and R. Yan. Merging storyboard strategies and automatic retrieval for improving interactive video search. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'07)*, pages 486–493, 2007.

[11] O. de Rooij, C. G. M. Snoek, and M. Worring. Query on demand video browsing. In *Proceedings of the ACM International Conference on Multimedia (ACM MM'07)*, pages 811–814, 2007.

[12] O. de Rooij, C. G. M. Snoek, and M. Worring. Balancing thread based navigation for targeted video search. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'08)*, pages 485–494, 2008.

[13] O. de Rooij and M. Worring. Browsing video along multiple threads. *IEEE Transactions on Multimedia*, 12(2):121–130, 2010.

[14] A. Hampapur, A. Gupta, B. Horowitz, C.-F. Shu, C. Fuller, J. R. Bach, M. Gorkani, and R. Jain. Virage video engine. In *Proceedings of the SPIE International Conference on Storage and Retrieval for Image and Video Databases*, pages 188–198, 1997.

[15] 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 *Proceedings of the ACM International Conference on Multimedia (ACM MM'06)*, pages 385–394, 2006.

[16] D. Heesch. A survey of browsing models for content-based image retrieval. *Multimedia Tools and Applications*, 40(2):261–284, 2008.

[17] D. Heesch, A. Yavlinsky, and S. M. Rüger NN^k networks and automated annotation for browsing large image collections from the world wide web. In *Proceedings of the ACM International Conference on Multimedia (ACM MM'06)*, pages 493–494, 2006.

[18] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: A review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[19] B. Moghaddam, Q. Tian, N. Lesh, C. Shen and T. Huang. Visualization and user-modeling for browsing personal photo libraries. *International Journal of Computer Vision*, 56(1-2):109–130, 2004.

[20] G. Nguyen and M. Worring. Interactive access to large image collections using similarity-based visualization. *Journal of Visual Languages and Computing*, 19(2):203–224, 2008.

[21] M. Rautiainen, T. Ojala, and T. Seppänen. Cluster-temporal browsing of large news video databases. In *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME'04)*, pages 751–754, 2004.

[22] A. Rocha, J. Almeida, M. A. Nascimento, R. Torres, and S. Goldenstein. Efficient and flexible cluster-and-search approach for cbir. In *Proceedings of the International Conference on Advanced Concepts for Intelligent Vision Systems (ACIVS'08)*, pages 77–88, 2008.

[23] C. G. M. Snoek, M. Worring, D. Koelma, and A. W. M. Smeulders. A learned lexicon-driven paradigm for interactive video retrieval. *IEEE Transactions on Multimedia*, 9(2):280–292, 2007.

[24] M. J. Swain and B. H. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[25] M. Worring, C. G. M. Snoek, O. de Rooij, G. P. Nguyen, and A. W. M. Smeulders. The *mediamill* semantic video search engine. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'07)*, pages 1213–1216, 2007.

[26] M. Worring, C. G. M. Snoek, D. Koelma, G. P. Nguyen, and O. de Rooij. Lexicon-based browsers for searching in news video archives. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR'06)*, pages 1256–1259, 2006.

[27] E. Zavesky and S.-F. Chang. CuZero: embracing the frontier of interactive visual search for informed users. In *Proceedings of the ACM International Workshop on Multimedia Information Retrieval (MIR'08)*, pages 237–244, 2008.

[28] E. Zavesky, S.-F. Chang, and C.-C. Yang. Visual islands: intuitive browsing of visual search results. In *Proceedings of the ACM International Conference on Image and Video Retrieval (CIVR'08)*, pages 617–626, 2008.