A Filtering Tool to Support Interactive Search in Internet Video Archives

Manfred Del Fabro, Klaus Schoeffmann, Mario Guggenberger, Mario Taschwer
ITEC - Information Technology
Alpen-Adria-Universität Klagenfurt
Klagenfurt, Austria
{manfred,ks,mg,mt}@itec.aau.at

Abstract—We present a novel interface for large-scale video archives that uses content-based filtering of search results. The interface has been used for the Interactive Known-Item Search (Interactive KIS) task of TRECVID 2012 and achieved good search performance. We found that for KIS tasks content-based filtering as used in our interface is convenient and able to successfully narrow down interactive search for many of the TRECVID KIS queries used in 2012. We demonstrate the usefulness of our approach, where three types of content-based filters (color detection, audio classification and video concept detection) can be combined.

I. INTRODUCTION

When searching for popular topics in Internet video archives like YouTube, the result set often contains too many videos showing similar content. If a user is looking for a video that is not similar to the first videos in the results and even not related to them, the search often degenerates to a very tedious visual filtering process. Therefore, additional possibilities are needed for users to specify better what they are searching for.

We present an interactive tool for quick filtering of Internet video search results based on color analysis, music detection, video concept detection and textual filtering. By applying some simple filter options, it is possible to reduce the size of the result set tremendously and thus it should be easier to identify the searched item.

We used this interface for the Interactive Known-Item Search task of TRECVID [1]. The videos used for the evaluation were downloaded from the Internet Archive and show similar properties than YouTube videos with respect to the available metadata (title, description and tags) as well as the length of the videos. The task simulates an Internet video search in nature, but not in scale. The test set only contains about 8000 videos. More details about the TRECVID 2012 Known-Item Search task as well as the evaluation results can be obtained from the proceedings [2].

II. RELATED WORK

Cunningham and Nichols [3] present a study about how people find videos on the Internet. They identify four major actions that are performed by users: (i) search by using queries, (ii) browsing the result set, (iii) viewing the videos, and (iv) reading text associated with videos. The results show that approximately two thirds of the tasks started with searching. However, after the initial search the predominantly performed actions were browsing and viewing. This is a clear indication that user interfaces like the one presented in this paper are desirable to support users in their browsing activities after the initial search.

A summary of three studies about the search behavior of real world users for news video retrieval is presented by Christel [4]. The presented results show that both experts and novices achieved high performance in retrieval tasks, although relying on different capabilities like query-by-text, query-by-image and query-by-concept. Especially expert users prefer to have all three search capabilities. They perform significantly better having such a complete system instead of having only query-by-text capabilities.

Rode et al. [5] also performed user studies in interactive video search where they compared the performance of non-expert and expert users. The results of experts show less performance fluctuations than the ones of non-experts, while it can be stated that experts used query-by-concept search more frequently than non-experts. Paying attention to the problem that non-experts have more difficulties in using advanced search features Hopfgartner et al. [6] introduce an interface for browsing user generated videos on a tablet. A user study shows that with the help of that tool novice users are able to achieve a similar search performance like expert users.

An evaluation of content-based filters for image and video retrieval is presented by Christel et al. [7]. They investigated different distributions of the automatic classification accuracy of visual classifiers with respect to interactive filtering tasks. They propose a confidence distribution that should at least be achieved to serve as base for filtering approaches. Furthermore, it is stated that feature filtering makes only sense for large result set sizes.

During the three years (2010-2012), where the TRECVID Known-Item Search task was performed, several interactive approaches for tackling that specific problem were proposed and compared with each other. Details about all presented tools are given in the TRECVID proceedings [2], [8], [9].

III. FILTERING OF LARGE RESULT SETS

Our tool was used for quick interactive filtering of our automatically obtained results for the TRECVID 2012 Known-Item Search task. The system used for the automatic retrieval task is described in our TRECVID report [10]. Here we focus only on aspects related to the interactive filtering of the results.

In Figure 1 a screenshot of our filtering tool is shown. The search area at the top of the screen (Search) can be used to
indicate specific terms that should be contained in the metadata of the searched video, regardless of the initial query. Negation of terms is also possible as well as filtering the search results by language.

On the left hand side color filters (Color) can be applied to refine the results based on the detection of these colors in the video. By selecting a color (+) only those videos remain in the result set where our color detection algorithm detected a prominent occurrence of that color. Otherwise, if a color is deselected (-), all videos containing segments in that color are removed from the result set. On the right hand side the result set can be filtered by concepts (Concepts). The selection and deselection of concepts works the same way like for colors. For example, if an indoor scene is searched, the concept Landscape may be deselected. Our audio classification detects two types of audio segments: (i) speech and (ii) music.

A. Color Detection

We developed an own color detection algorithm to enhance the textual metadata of each video with color names based on color analysis of the videos’ key-frames. For that purpose the HSV color representation \((H..Hue[0..360], S..Saturation[0..100], V..Value[0..100])\) of an image is segmented into eleven channels using the following rules on a per pixel basis:

- **White/Bright**: \(V \geq 70, S < 10\)
- **Gray**: \(V \geq 50, S < 15\)
- **Black/Dark**: \(V \leq 30\) or \(V \leq 40, S \leq 5\)
- **Brown**: \((H \leq 8 \text{ or } H \geq 340)\) and \((15 \leq S \leq 50, 15 \leq V \leq 50)\)
- **Red**: \((H \leq 8 \text{ or } H \geq 340)\) and \(V > 15, S > 15\) and \((S > 50 \text{ or } V > 50)\)
- **Orange**: \((H \geq 8 \text{ or } H \leq 40)\) and \(V > 15, S > 15\) and \((S \geq 70 \text{ or } V \geq 90)\)
- **Yellow**: \(H \geq 40, H \leq 62\) and \(V > 15, S > 15\)
- **Green**: \(H > 62, H \leq 180\) and \(V > 15, S > 15\)
- **Cyan**: \(H > 170, H \leq 200\) and \(V > 15, S > 15\)
- **Blue**: \(H > 200, H \leq 258\) and \(V > 15, S > 15\)
- **Pink**: \(H > 258, H \leq 350\) and \(V > 15, S > 15\)

In order to account for slight variations of the corresponding color the Hue ranges for the different channels are rather wide and tolerant. The exact thresholds for these rules have been determined by empirical investigations with key-frames of the TRECVID 2012 dataset. If any key-frame of a video contains a color channel covering at least 30% of pixels, the corresponding color name is added as additional metadata to the video.

B. Concept Detection

For the concept filtering, we used the results provided by participants of the TRECVID 2012 Semantic Indexing (SIN) task, which had to be performed on the same dataset. 91 runs were submitted by 28 teams. For each of 19 predefined concepts 2000 video shots were returned by each team. A heuristic voting scheme was applied to select only those concepts for a video that occur in many runs and only at top ranked positions in the result set.

For concept \(c\) and video shot \(s\), let \(n(c,s)\) be the number of runs containing the given (concept, video shot) pair, and let \(r(c,s)\) be the average rank in these runs. We then selected pair \((c,s)\) only if both of the following conditions were met:

\[
\begin{align*}
n(c,s) & \geq T_n \\
\frac{C(c)}{r(c,s)} & \geq T_m
\end{align*}
\]

where \(C(c)\) denotes a prior confidence score of concept \(c\), and \(T_n\) and \(T_m\) are fixed thresholds. Prior confidence scores should accommodate the fact that some concepts are harder to detect than others, leading to less reliable retrieval results in all runs. All of these parameters were adapted heuristically by manual inspection of generated annotations on the TRECVID 2011 dataset. This dataset consists of a large number (\(> 8000\)) of diverse Internet videos. Therefore, the presented methods should be applicable to any Internet video archive.
C. Music & Speech Detection

To further enrich the metadata with information on whether or not music is part of the videos’ (optional) audio tracks, we implemented the Continuous Frequency Activation feature proposed by Seyerlehner et al. [11]. It was originally developed on demand of a television broadcasting company to determine where in the soundtrack of a TV production music is being played, and basically works by looking for stationary parts in the audio spectrogram. The big advantage of the feature is that it can detect music even if it is played in the background at low volume with louder signals like speech or noise being present. The resulting feature vector expresses which segments of a stream contain music. We add a music flag to the metadata of a video if at least half of that video is underlayed with music.

For the speech detection a simple approach was applied. TRECVID provided text transcripts of all audio tracks in the test set. Whenever text transcripts were available for a video, we additionally annotated it with a speech flag.

D. Interactive Search

Instead of the typically used list or grid view, we present the resulting videos one-by-one, but show more details for every video by showing key-frames from all shots. All key-frames are put together to form one single image, which provides a good overview of the video content at a glance. Figure 2 displays an example of such a composite frame. Even if a video consists of many shots, it is still possible to identify details in the composite frame. Typically, the number of shots in Internet videos is not too high, because such videos have rather short duration. A study showed that YouTube videos typically have a length of about 4 minutes on average. The videos in the test set used for our study have a duration between 10 seconds and 4 minutes.

The results of a video search task can be investigated by switching sequentially from one composite image to the other. The ordering of results is based on the rank of the videos after the initial query. For every query the 800 best-matching results are returned, which corresponds approximately to 10% of the video archive. The amount of videos in the result set can be reduced by applying the above described filter options. If a user wants to examine the video itself and not only composite images, it is possible to open a separate video window by clicking on the composite frame. A screenshot of the video window is shown in Figure 3.

IV. TRECVID 2012 PERFORMANCE

We performed an evaluation of the presented interface for the TRECVID 2012 Known-Item Search task. All together, the participants of our study had to carry out 48 search tasks. In 29 cases they were able to find the searched video within a time period of 5 minutes. In 19 cases the searched video was not found. This search performance led to the third place in the Interactive Known-Item Search ranking of TRECVID 2012 [2].

The main reason why some videos were not found is that in 12 cases the searched video was not included in the first 800 results of our automatic retrieval approach. Therefore, the users participating in our study were simply not able to find them. Paying attention to this fact, the users only missed 7 out of 36 videos. The average search time per query was 2 minutes and 41 seconds.

V. Evaluation

Regarding the methods applied for searching the videos, it seems that people are used to formulating text queries in interactive search tasks. As Table I shows, our users heavily relied on the text filtering feature, other filters (language, colors, concepts, and audio) were hardly used.

After the TRECVID evaluation we investigated the potential benefit of filtering search results. This evaluation revealed that our users in the TRECVID 2012 KIS task would have been much faster if they had used the filters more often. For that purpose, we performed experiments to estimate the potential gain of using the filter options of our interactive video search tool. The results of this experiment are shown in Table II. For each search topic (ID - the original TRECVID IDs are used), the result set was filtered using each color, each concept and all possible combinations of one concept and one color. We compare the rank of the searched video in the original result set (Search), based on the textual query, with the new

---

rank (Filtered) of the searched video after applying the best filtering option (Filters). Column Diff expresses the gain (in number of positions) that was achieved by the filtering. Where no positions are given, the searched video was not included in the first 800 search results, thus it could not be found with our interactive filtering tool.

We found examples in the test set where the searched video can be found quickly, although it is initially not even ranked in the top 100 results. Just by applying the presented filter options it can be achieved that many of the searched videos show up in the top 10 results. In four cases (898, 900, 901, and 903) the usage of only one color filter leads to the best possible rank in the result set. In two cases (894 and 905) the usage of a single concept filter results in the best rank. We identified another four search topics (891, 892, 906, and 908) where one concept and one color filter have to be combined to reach the best result. In six cases (895, 896, 897, 904, 912, and 914) the automatic retrieval algorithm lists the searched item already on the first position of the result set and in another case (907) the automatically achieved result could not be improved by applying any of the filters.

A further investigation was performed to evaluate the audio-based filtering. We extended the above described experiment by additionally applying the music and the speech filter to each filtering combination used before. For the test set and queries used in our evaluation, the impact of the audio-based filters is rather low. The speech filter improved the rank of the searched video for topic 900 by one position. The music filter improved the result of topic 903 by three positions. In all other cases the audio-based filter options did not lead to a better rank of the searched video.

VI. Conclusion

We presented a tool for interactive filtering of large result sets in Internet video search, which was used for the evaluation of the TRECVID 2012 Known-Item Search task. It can be applied to independent datasets if in addition to our color and music detection, also own concept detection and speech recognition algorithms were applied. For the experiments shown in this paper we relied on data provided by TRECVID for the latter.

The participants of our study heavily relied on textual search. Additionally performed experiments show that by applying only one or two filter options the search performance may be considerably improved. Therefore, we are going to perform further work regarding that topic to develop interaction concepts that animate users to use such filtering options in interactive video search. One important aspect, which has to be investigated in that context, is finding out how the best suited filters can be automatically suggested to users based on their initial query.

<table>
<thead>
<tr>
<th>Filter</th>
<th>Avg Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>8.33</td>
</tr>
<tr>
<td>Language</td>
<td>0.04</td>
</tr>
<tr>
<td>Color</td>
<td>0.25</td>
</tr>
<tr>
<td>Concept</td>
<td>0.46</td>
</tr>
<tr>
<td>Audio</td>
<td>0.04</td>
</tr>
</tbody>
</table>

TABLE I. FILTERS APPLIED PER QUERY ON AVERAGE

<table>
<thead>
<tr>
<th>ID</th>
<th>Search</th>
<th>Filtered</th>
<th>Diff</th>
<th>Filters</th>
</tr>
</thead>
<tbody>
<tr>
<td>891</td>
<td>105</td>
<td>4</td>
<td>101</td>
<td>Landscape, Yellow</td>
</tr>
<tr>
<td>892</td>
<td>126</td>
<td>3</td>
<td>123</td>
<td>Bike/Board, Black (Gray)</td>
</tr>
<tr>
<td>893</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>894</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>Landscape</td>
</tr>
<tr>
<td>895</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(none)</td>
</tr>
<tr>
<td>896</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(none)</td>
</tr>
<tr>
<td>897</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(none)</td>
</tr>
<tr>
<td>898</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>Black</td>
</tr>
<tr>
<td>899</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>900</td>
<td>498</td>
<td>126</td>
<td>372</td>
<td>Yellow</td>
</tr>
<tr>
<td>901</td>
<td>13</td>
<td>1</td>
<td>12</td>
<td>Yellow</td>
</tr>
<tr>
<td>902</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>903</td>
<td>16</td>
<td>7</td>
<td>9</td>
<td>Gray</td>
</tr>
<tr>
<td>904</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(none)</td>
</tr>
<tr>
<td>905</td>
<td>28</td>
<td>2</td>
<td>26</td>
<td>Nighttime</td>
</tr>
<tr>
<td>906</td>
<td>97</td>
<td>29</td>
<td>68</td>
<td>Woman, Blue</td>
</tr>
<tr>
<td>907</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>(none)</td>
</tr>
<tr>
<td>908</td>
<td>133</td>
<td>1</td>
<td>132</td>
<td>Musician, Blue</td>
</tr>
<tr>
<td>909</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>910</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>911</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>912</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(none)</td>
</tr>
<tr>
<td>913</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>914</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>(none)</td>
</tr>
</tbody>
</table>

TABLE II. Comparison of original and filtered position in the result set

REFERENCES


Supported by Lakesides Labs, Klagenfurt, Austria and the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF-20214-22573-33955).