ITEC-UNIKLU
Known-Item Search Submission 2012

Manfred Del Fabro, Mathias Lux, Klaus Schoeffmann, Mario Taschwer
ITEC - Information Technology, Klagenfurt University
Klagenfurt, Austria
{manfred,mlux,ks,mt}@itec.aau.at

Oct 29, 2012

Abstract

In this report we describe our approach to the known-item search task for TRECVID 2012. We describe how we index available metadata and how we gain additional information about the videos using content-based analysis. A rule-based query expansion and query reduction method is applied to increase the number of relevant videos in automatic runs. Furthermore, we describe an approach for quick, interactive filtering of large result sets. We outline how the parameters of our system were tuned for the IACC dataset and discuss our TRECVID 2012 KIS results.

1 Introduction

This report outlines our system for the TRECVID 2012 known-item search task (KIS) and describes the results, which were achieved in the automatic and interactive runs of the benchmark. We were provided with 8263 videos including annotations, a master shot reference, automatic speech transcripts, and the submissions to the TRECVID 2012 semantic indexing task (SIN). For the benchmark a set of search topics was released, each of which describing a single known item (video) in the dataset. For the automatic search runs our task was to maximize the inverted rank of the correct result. Aim of the interactive search runs was to find a given video within a time period of five minutes.

We used the available metadata, extracted additional information using content-based analysis and the content-based information provided by TRECVID and indexed all videos of the dataset based on the extracted information using Apache Lucene.

The known-item search was split into two phases. First, the known-items of the search topics were searched with an automatic, text-based retrieval approach. The results have been submitted as automatic runs to the TRECVID KIS task. Then, we investigated an interactive search approach, which was based on the results of the automatic runs. The automatic search results in large result sets, in which it is still hard to find a searched item. We developed a simple, but efficient and effective graphical user interface, which assists users in filtering large result sets fast. Two evaluation runs of this approach have been submitted as interactive runs to the TRECVID KIS task.

After this short introduction, we present an overview of our retrieval system. We outline what information was used to index the dataset, how we extracted it and how we tried to achieve better results by modifying the queries based on pre-defined rules. Furthermore, we explain the parameter tuning of our system based on the TRECVID 2011 dataset and ground truth. An extensive part of this paper describes our interactive approach. Finally, we present and discuss the results achieved with the automatic and interactive results.

2 Known-Item Search Approach

In this year's known-item search task we completely relied on text-based retrieval methods. At our first attempt in TRECVID 2010[2] we additionally used content-based methods, but we had to recognize that the IACC dataset is too diverse to be able to achieve useful results with content-based methods. We did not participate in TRECVID 2011.

For the TRECVID 2012 known-item search task all videos were indexed based on the metadata contained in the dataset. In order to achieve better results the index was extended with additional textual information that we gained based on content-based analysis of the videos. The following methods were applied: concept detection, color detection, music detection, automatic speech recognition, and automatic translation.

Furthermore, to be able to cope with possible ambiguities of query terms a rule-based query expansion and query reduction method was used to increase the number of relevant results.

2.1 Extraction of Additional Data

In this Section we describe the content-based analysis methods, which were used to enrich the textual information in our search index.

2.1.1 Concept Detection

To enhance our text index with content-based concepts, we automatically annotated videos using the provided run submissions of the "light" Semantic Indexing (SIN) task. There were 91 runs submitted by 28 research groups, each containing 2000 retrieved video shots for each of 19 concepts (see Table 2).

To obtain annotations of high confidence, we employed a heuristic voting scheme to select only (concept, video shot) pairs that occur in many runs and at low (top) rank. 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:

$$n(c,s) \geq T_n \tag{1}$$

$$C(c) \cdot \frac{n(c,s)}{r(c,s)} \geq T_m \tag{2}$$

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. The eventually used parameter values are shown in Tables 1 and 2. Note that the prior confidence score of *Throwing* was set to 0 to effectively eliminate this concept, because SIN submissions appeared to be very unreliable on it.

Table 1: Parameters and mappings of heuristic voting scheme. (a) Thresholds for heuristic voting. (b) Mutually excluding concepts. (c) Concept expansion mapping.

(a)

T_n	$= 20$
T_m	$= 0.03$

(b)

Airplane_Flying	Boat_Ship
Male_Person	Female_Person

(c)

concept	expanded concepts
Infants	Child
Bicycling	Bicycling, Skateboard
Sitting_Down	Sitting
Swimming	Swimming, Water

Table 2: Concepts contained in the provided run submissions of the "light" Semantic Indexing task, assigned prior confidences, and number of videos automatically annotated with these concepts in the TRECVID 2012 dataset (total 5151 annotations of 3499 videos).

concept	confidence	#videos
Landscape	1.0	568
Male_Person	1.0	536
Computer_Or_Television_Screens	1.0	507
Female_Person	1.0	467
Scene_Text	1.0	437
Nighttime	1.0	381
Computers	1.0	355
Sitting_Down	0.6	321
Walking_Running	0.5	283
Airplane_Flying	0.6	244

concept	confidence	#videos
Boat_Ship	0.5	242
Singing	0.7	222
Infants	0.6	206
Stadium	0.7	198
Bicycling	0.6	195
Swimming	0.6	194
Instrumental_Musician	0.6	186
Indoor_Sports_Venue	0.5	116
Throwing	0.0	0

To further improve confidence of generated annotations, mutually excluding candidate concepts (see Table 1) for the same video shot were resolved by dropping the concept with lower metric (2). Moreover, some concepts were expanded to a list of more general concepts (see Table 1), as suggested by inspection of generated annotations on the TRECVID 2011 dataset. Finally, concept annotations of video shots were aggregated and indexed for entire videos only.

Automatic annotation of the TRECVID 2012 dataset assigned 5151 concept annotations to 3499 videos, i.e. to 42% of all 8263 videos. The numbers of annotated videos per concept are listed in Table 2. Note that these numbers were collected before concept expansion.

2.1.2 Color Detection

The textual metadata of every video is enhanced with color names based on color analysis of the videos' key-frames. For that purpose the HSV color representation ($H..Hue[0..359], S..Saturation[0..100], V..Value[0..100]$) of the image is segmented into 11 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**: ($H \leq 8$ or $H \geq 340$) and ($15 < S \leq 50, 15 < V \leq 50$)
- **Red**: ($H \leq 8$ or $H \geq 340$) and $V > 15, S > 15$ and ($S > 50$ or $V > 50$)
- **Orange**: ($H \geq 8$ or $H \leq 40$) and $V > 15, S > 15$ and ($S \geq 70$ 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 $V > 15, S > 15$

The exact thresholds for these rules have been determined by empirical investigations with key-frames of the TRECVID 2011 data set. 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.

2.2 Query Expansion and Query Reduction

We used both *query expansion* and *query reduction* to optimize the text queries. For the latter, we removed the following common "fillers" from the query text according to empirical investigations of the TRECVID KIS taks of 2011: `a, an, and, are, as, at, be, by, find, for, from, has, have, in, is, it, of, on, that, the, this, to, which, with`

For the query expansion, we used the rules given in Table 3 to expand each query. Most of these rules have been used in order to match the annotated concepts, described in Section 2.1.1.

3 Known-Item Search Parameter Tuning

We used TRECVID KIS 2011 data and ground truth for tuning the parameters of our automatic search engine. As a starting point we used a Lucene index generated from the basic metadata (.XML) provided for the TRECVID KIS 2011 data set (`index-2011`). In addition to that baseline index we used the annotation described in Section 2.1 (`s`=ASR metadata, `m`=music detection, `a`=concept annotation, `t`=metadata translation of , `c`=color annotation) as well as query expansion (`QE`) and query reduction (`QR`). Table 4 shows the achieved performance of the different settings sorted by mean inverted rank. The results show that we were able to increase the mean inverted rank from 0.2520 (baseline) up to 0.3042. We can also see that the best result was achieved without query expansion. However, the reason for that is that our query expansion method is optimized for 2012 data, as it contains many rules to expand the query with concept annotations (Section 2.1.1) that are only available for 2012. Hence, as setting for 2012 we used the following (which was also best for 2011 in terms of correct queries): `QE QR s m a t c`

4 Interactive Known-Item Search

Our interactive known-item search approach is based upon the results of the automatic runs. The idea is to provide a means for a quick investigation of a large result set to the users. Therefore, we present an interactive tool that can be used to filter large result sets based on different criteria. For the automatic search runs we had to submit the first 100 videos returned for a query. In our interactive approach we used the first 800 videos returned for a query. By applying different filter options users should be in a position to quickly sort out not relevant content and thus finding searched items faster.

In Figure 1 a screenshot of our proposed search interface is shown. It is an extension of our existing video browsing tool[1]. The window is divided into several areas, each of which offers certain filtering options. The search area at the top of the screen (*Search*) displays the search topics and the visual cues related to it. The textbox can be used to specify search terms that should be contained in the metadata of the searched video. Furthermore, it is possible to filter the search results by language. On the left side color filters can be applied to filter the result set based on the color detection. By selecting a color (+) only those videos remain in the result set that contain that color. Otherwise, if a color is unselected (-), all videos containing that color are removed from the result set.

On the right side of the window the result set can be filtered by detected concepts. The selection and unselection of concepts works in the same way like for colors. In the top right corner an audio based filtering can be performed. In this year's benchmark we offered only an option for filtering the result set based on detected music segments, but in future we also plan to incorporate a speech filter. The top left corner (*Topic*) is only needed for loading search topics and for displaying the remaining time.

Table 3: Query expansion terms (if query contains keyword it is expanded by the corresponding terms).

Keyword	**Expansion Terms**
no sound	noaudio
airplane	Airplane_Flying
babies	Infants baby
baby	Infants
ball	sports game
boat	Boat_Ship
boy	Male_Person man
car	street gray
danc	music
game	Walking_Running Stadium
girl	Female_Person woman
guitar	music Singing
hair	Male_Person Female_Person
her	Female_Person
his	Male_Person
man	Male_Person
man	boy
music	Singing song perform
play	Indoor music Indoor_Sports_Venue
purple	pink
rock	Singing music
screen	Scene_Text Computers Computer_Or_Television_Screens
sing	Singing
singing	music song perform
street	Walking_Running Bicycling
talk	Male_Person Female_Person
talking about	discuss speech talk
walk	Walking_Running Female_Person Male_Person
wear	Male_Person Female_Person
woman	Female_Person
woman	girl
glasses	Male_Person Female_Person
night	dark Nighttime
shirt	Male_Person Female_Person

Table 4: Performance of different indexes/parameters for 2011, sorted by mean inverted (MI) rank. `QE`=query expansion, `QR`=query reduction, `s`=ASR metadata, `m`=music detection, `a`=concept annotation, `t`=metadata translation of , `c`=color annotation

Setting	Correct Queries	Missed Queries	AVG Rank	MI Rank
index-2011 QE	167	224	11.3293	0.2441
index-2011	169	222	11.9882	0.2520
index-2011 QE QR	186	205	13.6559	0.2641
index-2011 QR	185	206	11.7405	0.2755
index-2011 s m a t c	183	208	11.3224	0.2783
index-2011 QE QR s m	186	205	12.2151	0.2817
index-2011 QE QR s m t	190	201	13.3790	0.2877
index-2011 QE s m a t c	182	209	11.1978	0.2881
index-2011 QE QR s m a	189	202	12.7407	0.2895
index-2011 QE QR s m a t	191	200	12.4869	0.2945
index-2011 QE QR s m a t c	193	198	12.7254	0.3020
index-2011 QR s m a t c	191	200	11.2565	0.3042

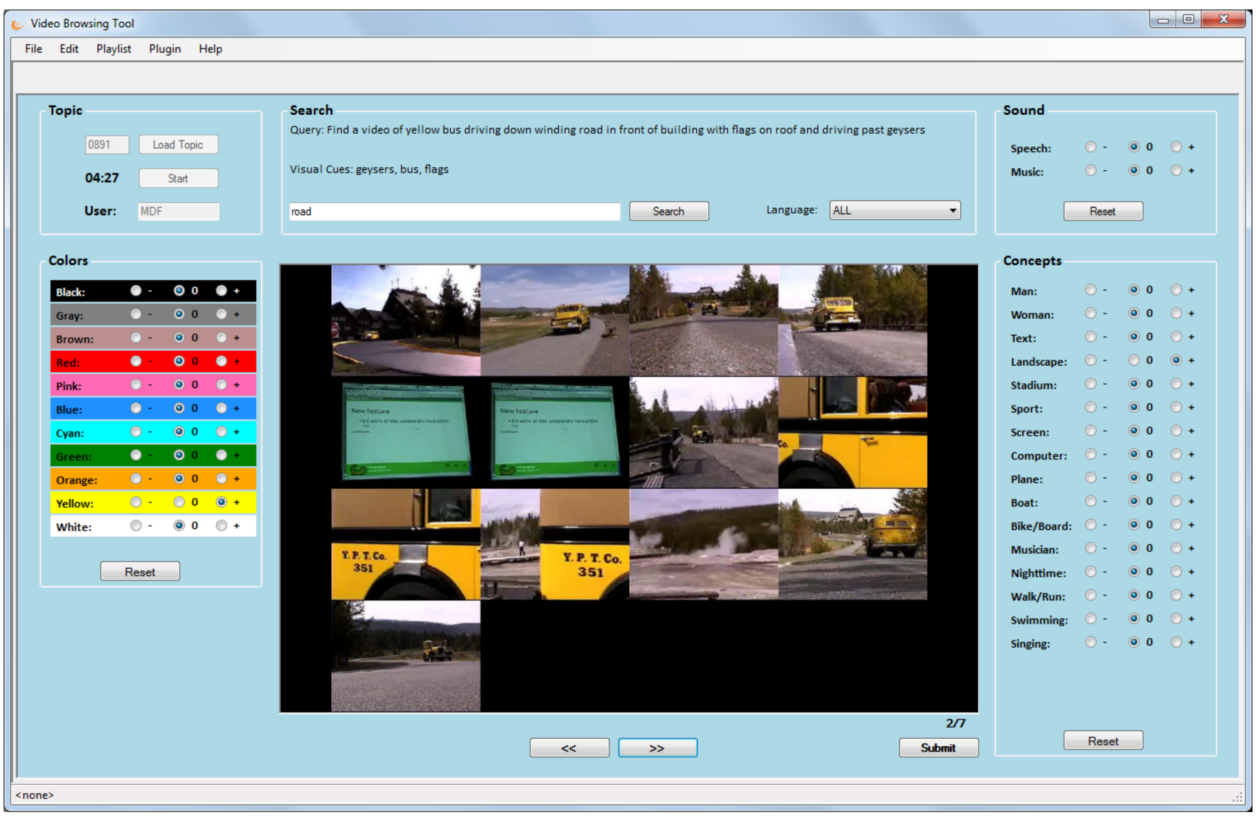

Figure 1: Screenshot of the interactive search GUI

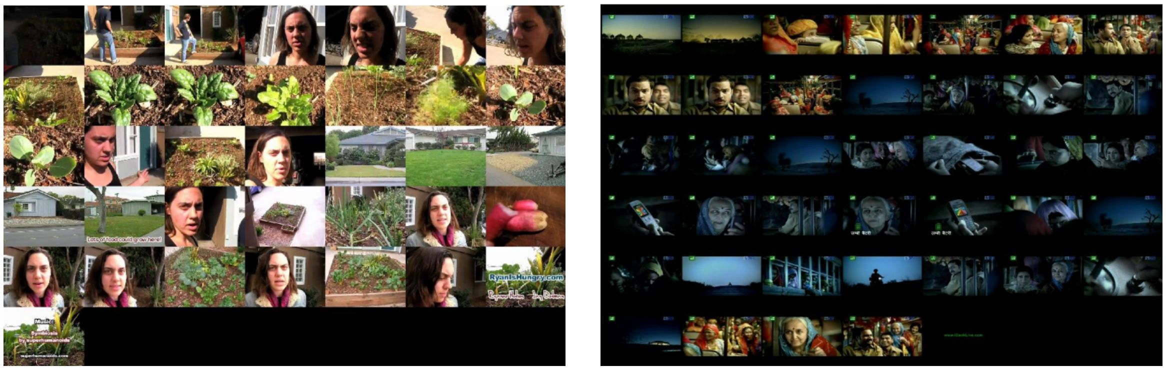

Figure 2: Examples for composite frames of two selected videos

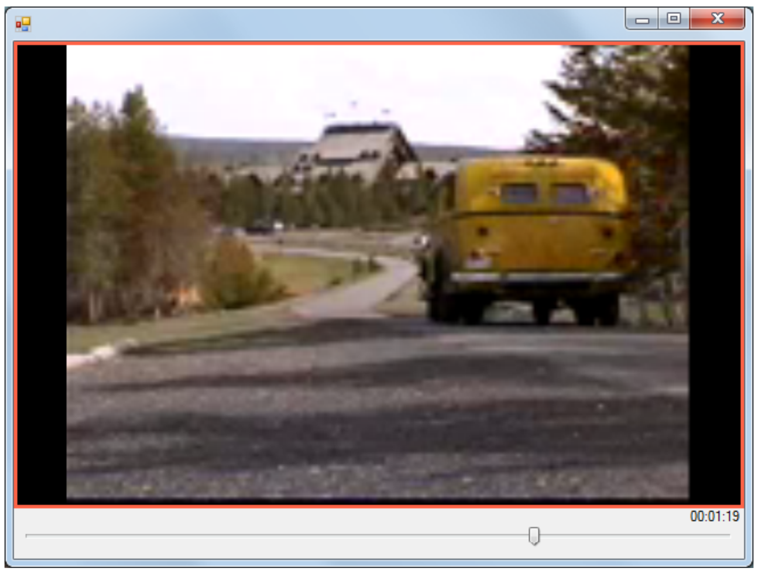

Figure 3: Screenshot of the video window

The results are not displayed in a list or grid view, which are commonly used views for presenting search results. Only one result is shown at a moment. Each video is represented by a composite image, which consists of all of its key-frames. We used the provided master shot reference and extracted the center frame of each shot as key-frame. By looking at composite images users can get an overview of the content of a video at a glance. The composite image in Figure 1 shows many details of the corresponding video. Even if a video consists of many shots, it is still possible identify a lot of details in the composite frame, as the two examples in Figure 2 show.

In our interactive approach, the result set is investigated by scanning from composite image to composite image. The ordering of results is based on the rank of the videos after the automatic run. The amount of videos in the result list can be reduced by applying the above described filters. If a user still wants to examine the video itself and not only by composite images, it is possible to open a separate video window by clicking on the composite image. A screenshot of the video window is shown in Figure 3.

With a small example we want to illustrate the usefulness of the proposed search interface and how it can basically be used to filter a large result set. For this purpose we use search topic 891: *find a video of yellow bus driving down winding road in front of building with flags on roof and driving past geysers*. Additionally, the video is characterized by the visual cues *geysers*, *bus*, and *flags*.

We applied different search options and investigated at which position the searched video is ranked in the remaining result set. The results are shown in Table 5. After the automatic run the searched video is only ranked at position 107, which means that it was not found in the automatic benchmark, because only the first 100 results are considered there. If we search for the terms indicated by the visual cues, we get a surprising result. For the given example the visual cues are not contained in the metadata of the video, thus they cannot be used to find it.

A closer look to the query shows that the term *road* may be useful, as well as the color *yellow* and the concept *landscape*. Filtering the result set only by the color yellow leads to rank 25, filtering only by the concept landscape leads to rank 18. This can even be improved by combining yellow and landscape, which leads to the fourth rank, without specifying any query term. If we search for the term road, the video shows up on rank 18. A combination with the concept leads to rank 6 and a combination with the color yellow results in rank 4. By combining all three filters the searched video even shows up in the second position. As this small example illustrates, by applying simple, but well thought filtering mechanisms it is possible to find searched items very fast, which have not been found at all during our automatic run at first.

Table 5: Filtering results for search topic 891 (t=used terms, cl=selected colors, cn=selected concepts)

Filter Options	Rank/Size Result Set
none (=result of automatic run)	107/800
t: bus	0/8
t: geysers	0/0
t: flag	0/11
cl: yellow	25/429
cn: landscape	18/85
t: road	18/82
cl: yellow + cn: landscape	4/52
t: road + cn: landscape	6/16
t: road + cl: yellow	4/18
t: road + cl: yellow + cn: landscape	2/7

5 ITEC-UNIKLU Known-item Search Results

5.1 Results of Automatic Runs

For the Automatic KIS task we used the setting with the best performance for 2011 (`QE QR s m a t c`, see Section 3). However, while we were able to achieve a mean inverted rank greater than 0.3 for 2011, in 2012 the same setting resulted in a mean inverted rank of 0.234 only (see Table 6, the setting used for the 2012 submission is marked with bold face). A post-hoc analysis with the ground truth data and different settings revealed that this result was even worse than with the baseline setting (`index-2012`). In other words, all the additional metadata (Section 2.1) as well as the query expansion made our results worse than with the baseline setting. The basic metadata (.XML) together with query reduction worked best with our system for the Automatic KIS 2012 task but this result (0.2409) nevertheless is very poor in comparison to all the other teams (see Fig. 4).

5.2 Results of Interactive Runs

As Figure 4 shows, we were able to reach promising results with the proposed interactive approach. Details about our two interactive runs are given in Table 7. The two participants of our study were able to find 15, respectively 14 videos out of 24 (*Success*).

On average 8 textual queries (*Query*) were used and 64 composite images (*Frames*) were investigated for a search topic. Additionally, more than 3 Videos were watched (*Video*) on average per topic. The mean search time for a query was about 161 seconds. In some cases our users were not able to identify the searched

Table 6: Performance of different indexes/parameters for 2012, sorted by mean inverted (MI) rank. `QE`=query expansion, `QR`=query reduction, `s`=ASR metadata, `m`=music detection, `a`=concept annotation, `t`=metadata translation of , `c`=color annotation

Setting	Correct Queries	Missed Queries	AVG Rank	MI Rank
index-2012 QE QR a	150	211	15.28	0.1842
index-2012 QE QR c	158	203	16.9241	0.191
index-2012 QE QR m	157	204	15.5032	0.1942
index-2012 QE QR t	156	205	14.5128	0.1949
index-2012 QE	155	206	12.6645	0.2182
index-2012 QE QR	164	197	11.9329	0.2238
index-2012 QE QR s	169	192	14.9172	0.2315
index-2012 QE QR s m a t c	164	197	12.4634	0.2342
index-2012	163	198	13.9141	0.2384
index-2012 s m a t c	163	198	11.0982	0.2398
index-2012 QR	168	193	13.0119	0.2409

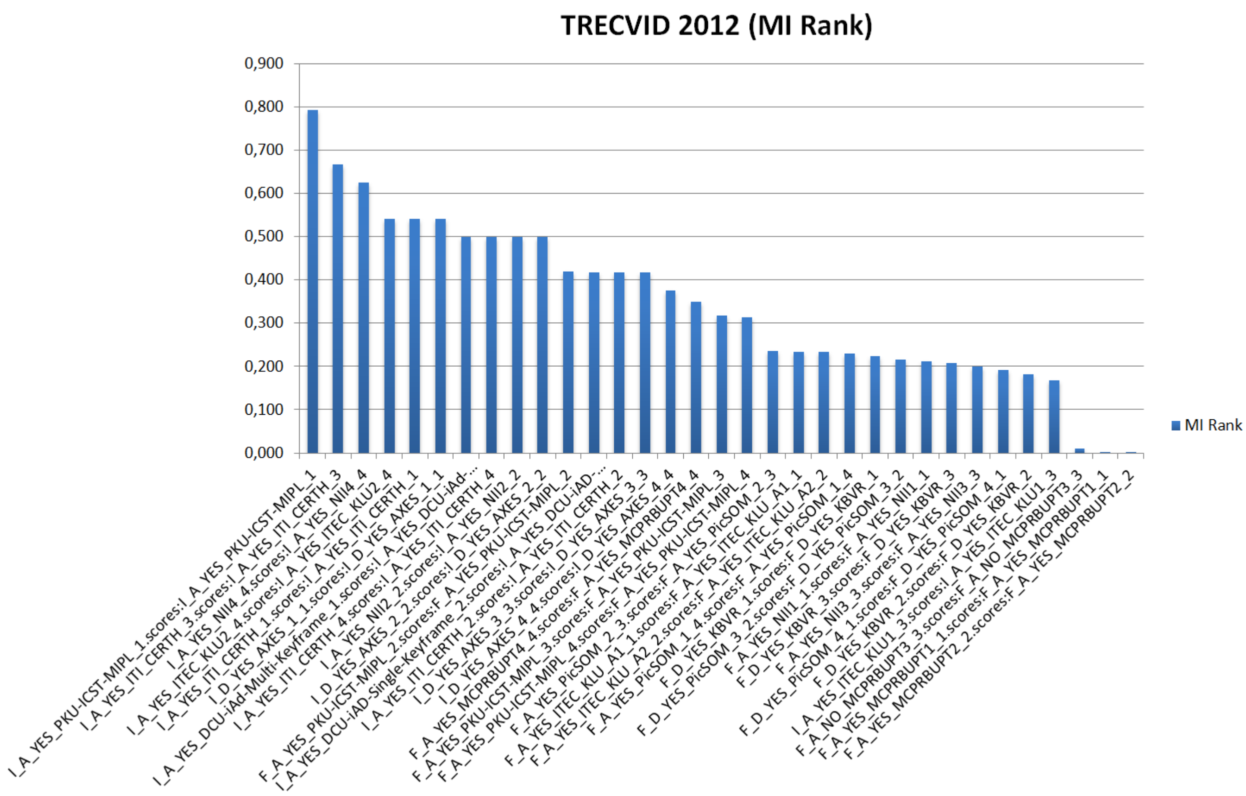

Figure 4: TRECVID Results 2012 for Automatic KIS

video immediately with the first submission and thus they had sent some wrong videos ids (*Wrong Subm.*) to the DCU oracle before the right video was found. In some cases our automatic runs were able to rank the searched videos already on the first positions in the result set (*Initial Rank*). Therefore, it was not difficult to find the corresponding videos. But on the other hand also some topics existed where the searched videos were ranked above position 100 (topics 891, 892, 900, 906, and 908). With the proposed GUI the users were still able to find those videos in reasonable time.

Overall, 8 videos have not been found at all. We took a closer look at the ranking produced by our automatic runs and we recognized that in 7 cases the searched videos were not contained in the first 800 automatic results. Therefore, it was not possible to find these videos with our proposed interactive approach. Interestingly, the content-based information that was additionally extracted from the videos (*Color*, *Concept*, *Music* and *Lang*) has hardly been used.

Table 7: Details of the submitted interactive runs (all values are averaged over both runs)

Topic	Query	Frames	Color	Concept	Music	Lang	Video	Wrong Subm.	Time (s)	Initial Rank	Success
891	10	64	1	0	0	0	3	0	122,5	107	true
892	14	102	0	1	0	0	7	4	251	130	true
893	16	121	0	1	0	1	5	1	300	0	false
894	0	2	0	0	0	0	0	0	15,5	3	true
895	0	0	0	0	0	0	1	0	7,5	1	true
896	2	0	0	0	0	0	1	0	15	1	true
897	0	0	0	0	0	0	1	0	9,5	1	true
898	1	1	0	0	0	0	1	0	148	5	true
899	15	98	1	2	0	0	5	0	300	0	false
900	10	62	0	3	0	0	13	2	282	513	true
901	18	184	1	0	1	0	4	2	300	18	false
902	14	148	0	0	0	0	4	1	300	0	false
903	1	6	0	0	0	0	1	0	44,5	16	true
904	0	0	0	0	0	0	1	0	10,5	1	true
905	15	117	0	0	0	0	5	0	194,5	28	true
906	1	2	0	0	0	0	1	0	98	99	true
907	13	107	0	0	0	0	7	1	175,5	2	true
908	1	8	0	0	0	0	1	0	65,5	141	true
909	9	106	0	0	0	0	9	1	300	0	false
910	19	133	0	0	0	0	5	0	300	0	false
911	16	124	3	1	0	0	3	0	300	0	false
912	0	0	0	0	0	0	1	0	7,5	1	true
913	25	155	0	3	0	0	1	0	300	0	false
914	0	0	0	0	0	0	1	0	12,5	1	true
Total	**200**	**1540**	**6**	**11**	**1**	**1**	**81**	**12**	**3859,50**		
Average	**8,33**	**64,17**	**0,25**	**0,46**	**0,04**	**0,04**	**3,38**	**0,50**	**160,81**		

6 Conclusion

In addition to the baseline metadata (given by the XML files of the videos) our system for the Automatic KIS task of TRECVID 2012 used metadata provided by automatic speech recognition, music detection, concept detection, translated XML metadata and color analysis. While this additional metadata worked well for the 2011 dataset and could improve the mean inverted rank, it led to worse results with the 2012 dataset. More precisely, our submitted result for the Automatic KIS task, which was based on all these metadata as well as query expansion and query reduction, achieved a worse mean inverted rank than the result based on baseline metadata only (0.2342 vs. 0.2384). It is not clear yet why all this additional effort/metadata could not provide a better rank and we will have to do a deeper analysis in order to find the reason for that.

On the other side we have seen that our approach for the Interactive KIS task works very well and ranked on 4th position in the overall results charts. However, as our analysis has shown, also in the Interactive KIS task we could hardly take advantage of the additional metadata at the competition although its usage would have significantly reduced the search time for some specific queries. Therefore, for TRECVID 2013 we need to investigate how we can better exploit all the additionally available metadata and achieve a better result than with baseline metadata only.

References

[1] Manfred del Fabro, Klaus Schoeffmann, and Laszlo Böszörmenyi. Instant video browsing: A tool for fast non-sequential hierarchical video browsing. In Gerhard Leitner, Martin Hitz, and Andreas Holzinger, editors, *HCI in Work and Learning, Life and Leisure*, volume 6389 of *Lecture Notes in Computer Science*, pages 443–446. Springer Berlin / Heidelberg, 2010.

[2] Mathias Lux, Klaus Schoeffmann, Manfred Del Fabro, Marian Kogler, and Mario Taschwer. ITEC-UNIKLU known-item search submission. In *Proceedings of TRECVID 2010*, pages 199–207. NIST, USA, 2010.