Interactive Evaluation of Video Browsing Tools

Werner Bailer[2], Klaus Schoeffmann[1], David Ahlström[1], Wolfgang Weiss[2], Manfred Del Fabro[1]

[1] Alpen-Adria-Universität Klagenfurt, Austria
ks@itec.aau.at,david.ahlstroem@aau.at,manfred@itec.aau.at
[2] DIGITAL – Institute for Information and Communication Technologies
JOANNEUM RESEARCH Forschungsgesellschaft mbH, Graz, Austria
{werner.bailer,wolfgang.weiss}@joanneum.at

Abstract. The Video Browser Showdown (VBS) is a live competition for evaluating video browsing tools regarding their efficiency at known-item search (KIS) tasks. The first VBS was held at MMM 2012 with eight teams working on 14 tasks, of which eight were completed by expert users and six by novices. We describe the details of the competition, analyze results regarding the performance of tools, the differences between the tasks and the nature of the false submissions.

1 Introduction

The Video Browser Showdown (VBS) is a live video browsing competition where international researchers, working in the field of interactive video search, evaluate and demonstrate the efficiency of their tools in presence of the audience. The aim of the VBS is to evaluate video browsing tools for efficiency at known-item search (KIS) tasks with a well-defined data set in direct comparison to other tools. For each task the moderator presents a target clip on a shared screen that is visible to all participants. The participants use their own computer to perform an interactive search in the specified video file taken from a common data set. The performance of participating tools is evaluated in terms of successful submissions and search time. The first VBS was held at the 18th International Conference on MultiMedia Modeling (MMM 2012) in Klagenfurt, Austria, where eight international teams participated (see [1] for descriptions of their systems). The setup of the room is shown in Figure 1. In this paper, we describe the details of the competition, analyze results and evaluate data collected during the competition for further insights.

2 Competition Details

The VBS consisted of 14 tasks. The first eight tasks were performed by experts (the developers of the tools) and the last six tasks were performed by eight volunteers randomly assigned to one of the tools. For each task the corresponding video containing a randomly selected target clip was mentioned first, in order to

Fig. 1. Photo of the VBS competition at MMM 2012 (credit: Rene Kaiser).

allow the participants to set up their systems for that video. After that, the 20 seconds long target clip was projected on a wall and the sound was played on loudspeakers. After the presentation, participants were given a maximum of two minutes to find the target clip. Participants were allowed to start their search already while the target clip was presented but could miss important information as the target clip did not necessarily contain redundant content. The collection from which videos for the tasks were randomly selected (see Table 1) consisted of 30 videos with an average length of 77 minutes (min: 31, max: 139).

Participants submitted found segments to an HTTP server with an URI that contained the following information: team, video, start and end frame number of the submitted segment. The server was responsible for three tasks: (1) checking whether the submitted segment was correct (i.e., overlapped with the target clip), (2) measuring the task solve time and (3) computing scores for all teams and tasks. In each run (expert and novice), the team with the highest sum of scores was the winner. A submitted segment was considered as correct if $(S_i - 125) \leq s_i \leq e_i \leq (E_i + 125)$, where s_i and e_i are the submitted start and end frame numbers and S_i and E_i are the begin and end frame numbers of the target clip for task i. For each task a maximum score of 100 could be obtained. The score was dependent on the task solve time (t_i) and a penalty based on the number of wrong submissions (w_i) for a task i before the correct submission was received: $score = (1/\max(1; w_i - 1))(100 - 50\frac{t_i}{T_{max}})$. T_{max} is the maximum time allowed for each task, i.e. 120 seconds. During the search, the current score, overall score, number of correct and wrong submission for all teams were projected on the wall.

Task	Video	Language	Genre	Duration
1	8	JP+EN	Documentation	01:15:01
2	11	NL	News	01:07:30
3	1	IT	Talkshow	01:50:02
4	26	IT	Talkshow	01:27:56
5	14	NL	News	01:05:06
6	9	IT	Talkshow	01:32:33
7	20	DE	News	00:55:32
8	13	NL	News	00:30:36
9	5	JP+EN	Documentation	01:15:01
10	20	DE	News	00:55:32
11	15	EN	News	00:55:59
12	29	JP+EN	Documentation	01:15:01
13	13	NL	News	00:30:36
14	28	JP+EN	Documentation	01:15:00

Table 1. Videos used in VBS (expert tasks at top, novice tasks at bottom).

3 Evaluation

In the following, we describe the use of the data gathered during the VBS to compare the performance of tools, the different tasks and analyze the false submissions.

3.1 Comparison of tools

Given the low number of completed tasks used in the VBS – eight for the experts and six for the novices – we limit our analyses to descriptive statistics. Figure 2 shows the combined team scores and scores from the expert and novice runs separately. Overall, the median score was 85.5 points (mean 65.7, s.d. 38.9). Team 1 had the best total score (expert + novice) of 1130 points and was closely followed by Team 2 and Team 8 with 1061 and 1048 points respectively. With scores between 853 and 933 points, there were close calls in the middle field between Team 3, 4, 5 and 6. Team 7 had the lowest score which was 55% lower than the winning Team 1 and 40% lower than the second last team, Team 6. A similar order among the teams is also visible in the data from the expert run. In the novice run the teams were positioned somewhat closer together with Team 5 and 8 having the highest scorers.

The box plot in Figure 3 provides a more detailed overview of how the different browsers performed. Team 1 was impressively consistent and scored 92 or above in 11 of the 14 tasks, resulting in a median score of 95.5. Task scores for the other teams were more varied, as indicated by the rather large interquartile ranges shown in Figure 3. Five of the expert users managed to score 100 points in at least one of the tasks and all of them failed in finding the correct segment in one or more tasks, and thus ended up with no points in these tasks. Generally,

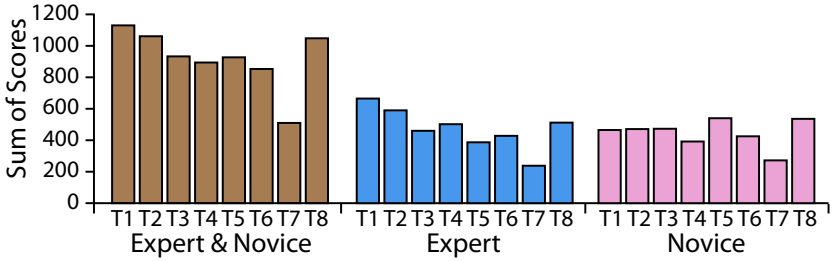

Fig. 2. Total team scores (T1 to T8) and scores in the expert run and novice run separately.

the novices were more consistent in their scoring than the experts were (however, note that experts performed two more tasks). All but one of the novices (Team 2) had a higher median score than their expert team colleague. Except from the novice user in Team 2, all novices scored full points in at least one task and three of them (Team 2, 5 and 8) were successful in all of their six tasks, achieving scores of 20 points or above.

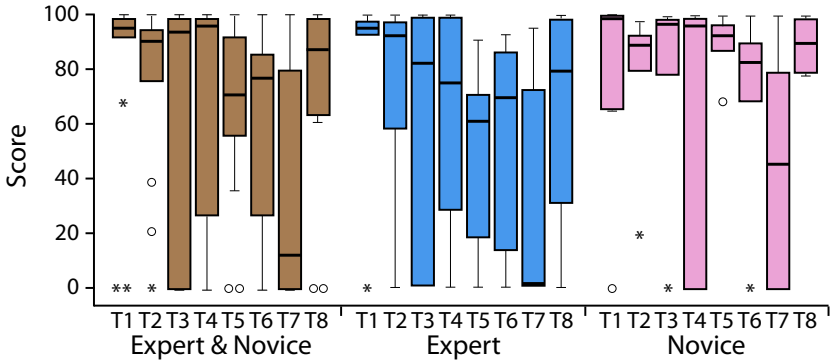

Fig. 3. Box plot of combined team scores (expert+novice), and separated by expert and novice (○: >1.5 IQR, ⋆: >3 IQR).

In summary, with the limited data collected, no inferential statistical procedure is applicable that let us draw any conclusions about performance differences between the browsers. However, the above analysis provides indications that some browser designs might have been more suitable than others for the video browsing tasks we used. In particular, we see a tendency of consistent and good performance of Team 1's browser. Similarly, the low scores for Team 7 might also indicate a sub-optimal design. A detailed description of these two browsers is provided in Section 4.

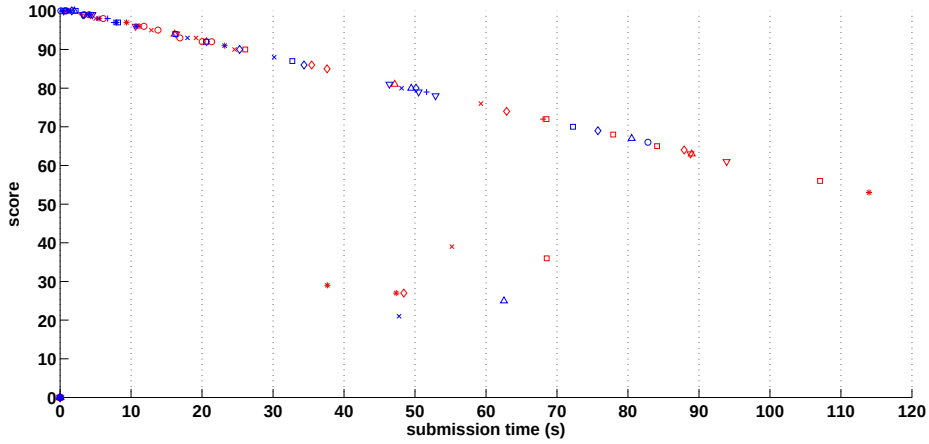

Fig. 4. Score vs. submission time of correct results (red = experts, blue = novices, the symbols denote the different teams).

Figure 4 shows the relation between submission time and score. It is apparent that the submission time is the most important component in the score, resulting in a linear dependency along the maximum score possible at a certain time into the task. Only 7 of the 88 correct submissions have reduced score due to prior false submission (5 from the expert run, and of those 2 from the same team, and 2 from the novice run). It is interesting that none of the false submissions after which a correct result was submitted, were submitted before 70s into the task time (i.e., mostly in the first half of the working time, while after that users seem to have a clear idea about the distinctive features of the target clip).

3.2 Comparison of tasks

Figure 5 shows the correct and false submissions for expert and novice runs. It is evident that novice users submitted a lower number of false results than expert users at a comparable number of correct ones. Furthermore, novice users made their correct submissions no later than 85s after trial start, whereas expert users often submitted both correct and false results until just before the allowed 120 seconds had elapsed. One interesting observation from these results is that the lines for correct and false submission in the figure rarely cross after the first 10-20 seconds, i.e., for the rest of the task duration, the number of false submissions always either stays below or above the number of correct submissions. Only for the first two tasks in the expert run several crossings occur, while in the other cases one can already predict the average success rate quite well after 20 seconds. As these were the first two tasks performed in the competition, we probably cannot draw any conclusions from this fact.

Can we draw conclusions regarding the difficulty of each task from the measured results? Looking at the scores per task (Figure 6), we can see that tasks

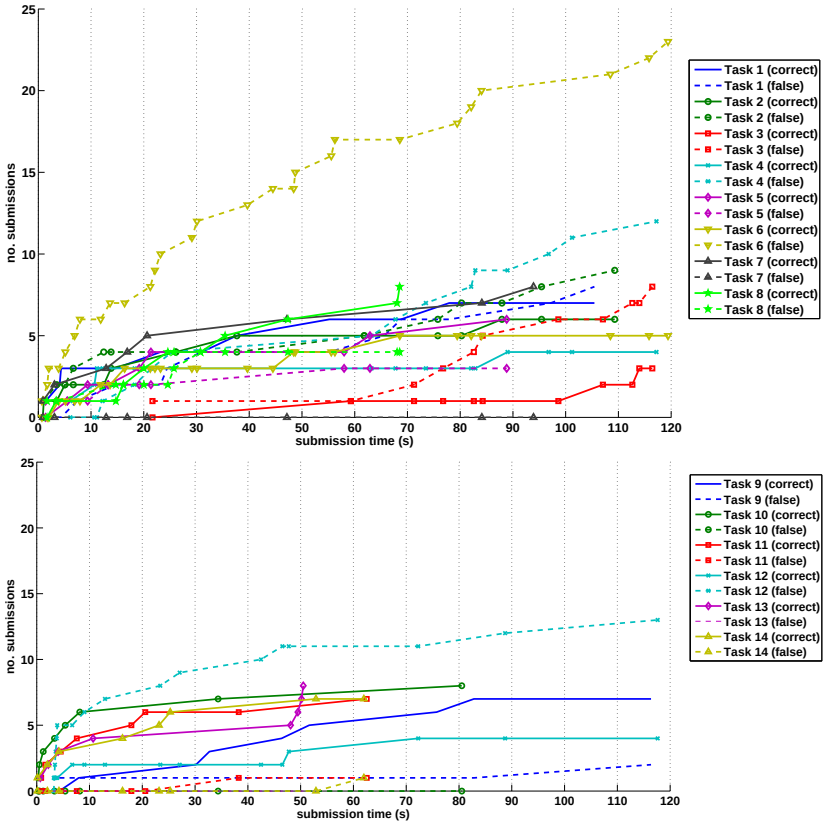

Fig. 5. Correct (solid) and false (dashed) submissions over time per task for expert (top) and novice runs (bottom).

3, 4, 6 and 12 have lower mean scores than all others, and that at least the lower quartile of scores is 0. We see a similar pattern in the submission times of the correct results (setting it to 120 seconds if no correct result has been submitted): Tasks 3, 4, and 12 have clearly higher mean duration (for task 6 it is only slightly higher than the others), but for all of them at least the upper quartile is at the maximum duration of the task. The false submissions show a slightly different picture: Tasks 4, 6 and 12 have the highest number of false submissions. However, task 3 has only one false submission, probably due to the fact that the video contains no other segment sufficiently similar to the target clip than the correct one. The contrary is true for tasks 2 and 8: both have a rather high number of false submissions, but the mean submission time is only 30 seconds, and more than 75% of the teams completed the task in time (some after an initial false submissions). It seems that the submission time and the score (which is linearly dependent on submission time for tasks with few false positives) are good indicators of tasks for which it is hard to locate a relevant

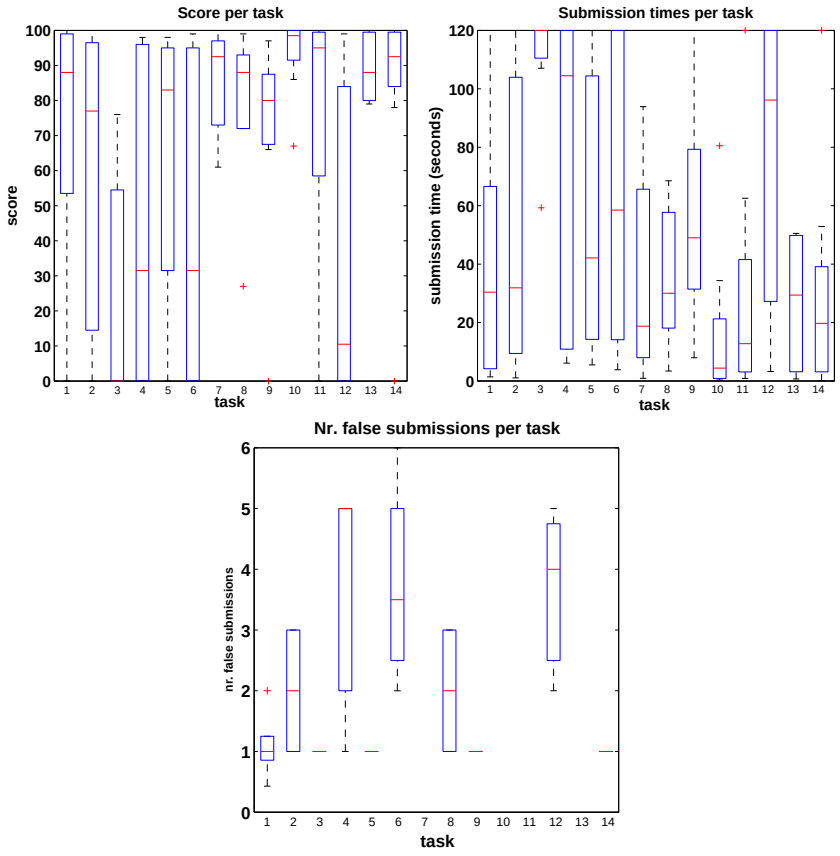

Fig. 6. Scores, submission time and misses per task.

segment. However, these measures do not capture the cases where a number of similar candidate segments increases the risk of making a false submission. Looking at the submissions over time in Figure 5, we see that the tasks that are salient in the plots in Figure 6 (task 3, 4, 6, 12) have consistently a higher number of false than correct submissions from about 20s into the task time to the end.

3.3 Analysis of false submissions

In total, 172 results were submitted for all the tasks, 88 were correct and 84 were false. In order to answer the question "how false are the false submissions" we manually examined each of the false submissions. It turns out that 82% of them are globally visually similar to the target clip (i.e., have a similar layout and colors), 38% have a similar background as the target clip, and 49% show a similar object as the target clip. This means that depending on the type of material

request, many of these results could already be useful results for a material search task in a media production scenario. Apart from visually similarity, a request could be to find a video clip about a certain topic. Here we see that 46% of the false positives are from the same scene or news story as the target clip, i.e., participants were close to correct.

We also analyzed the temporal distance between a false submission and the ground truth. It is smaller in news and documentaries (as they tend to be more structured) and larger in talk shows, where similar segments occur at different times. There are some exceptions to this trend. For example, the video used in task 6 is a talk show, but all segments that are visually similar to the query segment are close to the true result, resulting in an atypically low temporal distance of the false submissions. On the other hand, the video in task 11 is made of two news broadcasts, but there are several false submissions far from the true result. The reason is that the target clip is a phone interview, showing a map of Haiti, and many false positives are from an interview showing a similarly designed map of Sri Lanka.

4 System Analysis

In this section we present a top and a low scoring video browsing system that attended the showdown in order to investigate which are successful approaches for a KIS task. Details about the systems can be found in [1].

4.1 Top scoring video browser of Team 1

Interestingly, the AAU Video Browser (see Fig. 7), scoring best in the expert run, abstains from content analysis. Instead, intelligent interaction means for video browsing are combined with the human ability to identify relevant content very fast. During the competition a simple search strategy can be observed. A combination of parallel video browsing and hierarchical video browsing is applied to all tasks. First the parallel mode was used to divide each video into four or even nine equal parts to be shown in parallel. The decision to use four or nine video windows depends on type of the video. If a video contains a lot of scenes with similar content, only four parts were used, whereas for videos containing a lot of scenes with different content nine parallel windows were chosen. By dragging and pulling the timeline slider, all parts can be observed in parallel and thus it is possible to quickly get an overview of a video. If a candidate segment has been located, the hierarchical browsing mode can be used to investigate it in detail. This is done by recursively applying the parallel mode only to the candidate segment.

The VBS showed that this simple search strategy, combined with intelligent and easy to use interaction means, is very efficient for KIS tasks. In contrast to the human mind, an artificial concept classifier is only able to identify a limited amount of trained concepts. The dataset contained a broad range of different videos. It seems that the concept-based approaches are too limited regarding the

range of concepts they are able to identify in these videos. Therefore, not the most sophisticated tool was successful, but the one that supported the interactive search task best.

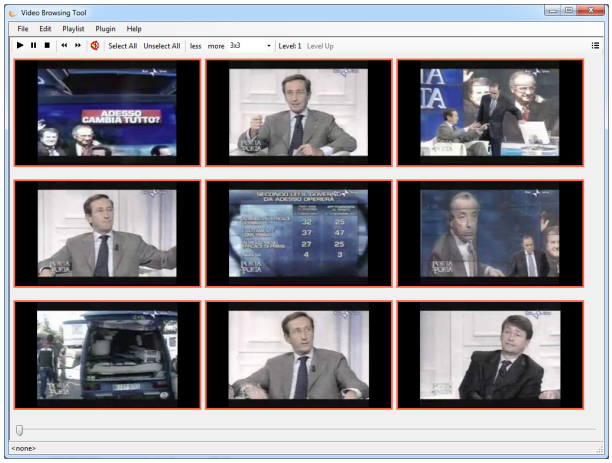

Fig. 7. Screenshot of the best scoring tool (AAU Video Browser).

4.2 Low scoring video browser of Team 7

Team 7 used a video browsing application targeted for content management in post-production of film and TV. It works based on automatic content analysis that performs camera motion estimation, visual activity estimation, extraction of global color features and estimation of object trajectories. The central component of the browsing tool's user interface is a light table that shows the current content set and cluster structure using a number of representative frames for each of the clusters (see also Fig. 8). The users apply an iterative content selection process, where they cluster content or search similar keyframes based on the extracted features and can then select relevant keyframes to reduce the content set.

Analysis on task level: The team ended at the bottom of the list with 510 points. The expert finished 3 out of 8 tasks successfully and needed on average 51 seconds for a successful task. The novice needed nearly the same amount of time of 49 seconds on average for a successful task and completed 4 out of 6 tasks successfully.

Analysis on submission level: At the expert run, the user made 16 false and 3 successful submissions. A wrong submission took in average 27 seconds where the user did about 10 interactions (e.g., clustering, similarity search, preview) on the user interface. In 94% of the cases of wrong submissions the user submitted a visually similar keyframe and 75% of the wrong submissions were in a scene of the

target clip. This means the user was temporally and visually near the searched item but still failed. A successful submission took on average 42 seconds with 19 interactions. The novice made 6 wrong and 4 successful submissions. In contrast to the expert, the novice submitted only in 17% of the wrong submissions a visually similar keyframe and 50% of the wrong submissions were in a scene of the target clip. A wrong submission took 32 seconds on average with 11.5 interactions and a successful submission took 38 seconds on average with 14.5 interactions.

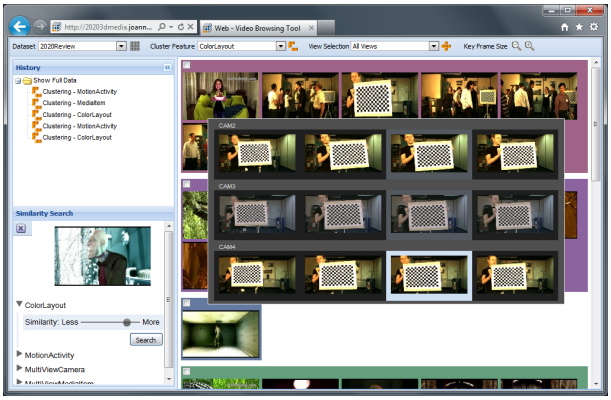

Fig. 8. Screenshot of the video browsing tool from team 7.

5 Conclusion

We have presented an analysis of the results of the Video Browser Showdown, an interactive evaluation of video browsing tools by solving known-item search (KIS) tasks, where eight teams worked on 14 tasks. The intent of the Video Browser Showdown is to demonstrate the efficiency of interactive video search tools for the real world scenario, where users want to quickly find content in videos, which is known to the user but cannot be found or effectively described by a query with a typical video retrieval tool. Because the competition addresses both expert and novice users, it should promote research on interactive video search tools that allow efficient search for the majority of users. It is an encouraging result that a correct keyframe was submitted before the allowed 120 seconds in 79% of the trials. Also, many of the false results were visually very similar and/or belonged to the same semantic unit of the content.

We have analyzed the tools of the best and the worst performing teams in order to get additional insigths about what approaches work and do not work for KIS tasks. It seems that for KIS tasks working with a small dataset, simplicity is the key. This observation holds for both expert and novice users,

which showed surprisingly little difference in their performance. The analysis of the different tasks shows that four of the fourteen tasks were particularly difficult, an observation that is supported by several of the measured parameters. Users tend to solve tasks with a first correct submission or not at all. The majority of all false submissions (72 out of 84) were registered in trials where no correct result has been submitted. Out of 88 correct submissions, 19 (22%) were made after a false submission, but only 7 (8%) received a reduced score due to penalties for multiple prior false submissions. As the scores were mainly related to the submission time, it could make sense to penalize also the first false submission in future competitions. It would also be interesting to compare different types of tasks, e.g., KIS on multi-item collections, as well as to consider related task settings motivated by problems in media production workflows. Furthermore, in order to allow for a deeper analysis and better understanding of how users interact with the system for particular tasks, systems participating in future events of this competition should implement an appropriate logging feature and provide log data for post-hoc analysis.

In summary, besides providing useful insights regarding the performance of various video browsing tools, the VBS competition did also allow participating teams to demonstrate their tools in a fun and exiting setting that was highly appreciated by other conference participants as a suitable round-up of the conference days.

Acknowledgments. The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 287532, "TOSCA-MP".

References

1. Klaus Schoeffmann, Bernard Mérialdo, Alexander G. Hauptmann, Chong-Wah Ngo, Yiannis Andreopoulos, and Christian Breiteneder, editors. *Advances in Multimedia Modeling - 18th International Conference, MMM 2012. Proceedings*. Springer, 2012.