Crowdsourcing Visual Detectors for Video Search

Bauke Freiburg
Video Dock
Panamalaan 1b, 1019 AS
Amsterdam, The Netherlands
bauke@videodock.nl

Jaap Kamps
ISLA, University of Amsterdam
Science Park 904, 1098 XH
Amsterdam, The Netherlands
kamps@uva.nl

Cees G.M. Snoek
ISLA, University of Amsterdam
Science Park 904, 1098 XH
Amsterdam, The Netherlands
cgmsnoek@uva.nl

ABSTRACT

In this paper, we study social tagging at the video fragment-level using a combination of automated content understanding and the wisdom of the crowds. We are interested in the question whether crowdsourcing can be beneficial to a video search engine that automatically recognizes video fragments on a semantic level. To answer this question, we perform a 3-month online field study with a concert video search engine targeted at a dedicated user-community of pop concert enthusiasts. We harvest the feedback of more than 500 active users and perform two experiments. In experiment 1 we measure user incentive to provide feedback, in experiment 2 we determine the tradeoff between feedback quality and quantity when aggregated over multiple users. Results show that users provide sufficient feedback, which becomes highly reliable when a crowd agreement of 67% is enforced.

Categories and Subject Descriptors: H.3.3 Information Storage and Retrieval: Information Search and Retrieval
General Terms: Algorithms, Experimentation, Performance
Keywords: Semantic indexing, video retrieval, information visualization

1. INTRODUCTION

Social tagging platforms like YouTube and Vimeo are effective for sharing complete videos, but users requiring access to video fragments have no other choice than to inefficiently browse through an entire video of interest [8]. For a long time, automated video-fragment classifiers promise to alleviate the manual burden of localizing specific video fragments. Despite good progress in video concept classification, the performance of automated methods still varies [9], which can be attributed to the amount of training data used. In this paper we study social tagging at the video-fragment level using a combination of automated content understanding and the wisdom of the crowds.

Area Chair: David Shamma.

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

We start from the state-of-the-art in concept detection [11], and we investigate how visual tagging by an online crowd might support video access at the fragment level. Crucial in such a crowdsourcing study is user tagging motivation. In-depth studies on motivations for visual tagging are many, emphasizing in particular organizational, social, and gaming aspects [5, 2, 1], or Fame, Fortune, Fun, and Fulfillment [6]. An additional motivation for (visual) tagging is the creation of new information. For example when incorrect individual tags are aggregated in the right way their collective judgement might be highly reliable [13]. In fact, a crowdsourcing evaluation can be as reliable as an expert evaluation [12]. However, it is well known that in online communities 90% of users never contributes [7]. To assure participation of the remaining 10% of users, micro payments are often offered [4], or the social impact is emphasized, especially within a dedicated user community. Surprisingly, exploiting the crowd for video labeling at the fragment level is scarce [8, 14, 3], to the best of our knowledge in combination with automated visual analysis non-existing even.

The main research question in this paper is: *can user tags from crowdsourcing be beneficial to a system that automatically predicts labels for video fragments?*. We divide this question into the following sub-questions: *i*) Does a concept-based video search engine provide enough incentives for users to provide labels, without receiving payment or other compensation? *ii*) Are the resulting labels of sufficient quality compared to expert labels, when aggregated over multiple users? In order to answer these question, we develop a video search engine for a dedicated user-community that allows for easy fragment-level crowdsourcing. We highlight the case study in which we evaluate our research next.

2. CROWDSOURCING CASE STUDY

We study the merit of crowdsourcing visual detectors for video search within a cultural heritage context. To maximize user participation, we motivate online users by providing them with access to a selection of exclusive, full-length concert videos. While our case study is limited to the cultural heritage domain by design, the principle of combining visual detectors with crowdsourcing is general and applicable to *any* multimedia retrieval, authoring, or visualization application open for cultural engagement.

2.1 Concert Video Search Engine

We developed an online video search engine for rock concerts [10]. Our search engine uses archived video footage of the Pinkpop festival. This annual rock festival is held

Figure 1: Eleven common concert concepts we detect automatically, and for which we collect user-feedback.

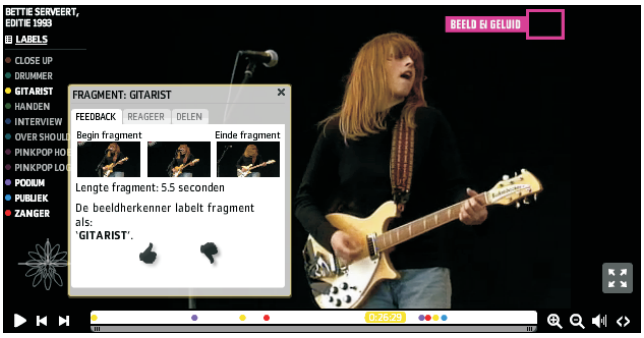

Figure 2: Timeline-based video player where colored dots correspond to automated visual detection results. Users can navigate directly to fragments of interest by interaction with the colored dots, which pop-up a feedback overlay as displayed in Figure 3.

since 1970 at Landgraaf, the Netherlands. All music videos have been recorded during the 40 years life cycle of the festival. We cleared copyright for several Dutch and Belgian artists playing at Pinkpop, including gigs from *K's Choice*, *Junkie XL*, and *Moke*. The amount of footage for each festival year varies from only a summary to almost unabridged concert recordings, even including raw, unpublished footage. The complete video archive contains 94 concerts covering 32 hours in total.

We create detectors for 11 concert concepts following a state-of-the-art implementation [10]. We select the concepts based on frequency, visual detection feasibility, previous mentioning in literature and expected utility for concert video users (summarized in Figure 1). We consider a video fragment a more user-friendly retrieval unit compared to more technically defined shots or keyframes. We create fragment-level detection scores from frame-level scores by aggregating the concept scores of all the frames in the processed videos. The fragment algorithm was designed to find the longest fragments with the highest average scores for a specific concert concept [10]. Users may provide feedback on these automatically detected fragments using our feedback mechanism.

Figure 3: Harvesting user feedback for video fragments (top to bottom). The thumbs-up button indicates agreement with the automatically detected label, thumbs-down disagreement. Three key frames represent the visual summary of the fragment. Users may correct wrong labels, adapt fragment boundaries, or suggest additional labels (in Dutch).

2.2 Feedback Mechanism

The main mode of user interaction with our video search engine is by means of the In-Video Browser, see Figure 2. The timeline-based browser enables users to watch and navigate through a single video concert. Little colored dots on the timeline mark the location of an interesting fragment corresponding to an automatically derived label. To inspect the label, users simply move their mouse cursor over the colored dot. By clicking on the dot, the player instantly starts the specific fragment in the video. If needed, the user can manually select more concept labels in the panel on the left of the video player. Too maintain overview, the In-Video browser automatically launches with a maximum of twelve fragments on the timeline interface every time a user starts a concert. These twelve correspond to the most reliable fragment labels. Once the timeline becomes too crowded as a result of multiple selected labels, the user may decide to zoom in on the timeline to retrieve fragments for a specific, smaller part of the video.

An important aspect of the In-Video browser is that the user viewing experience is interrupted as little as possible, the video continues to play while the user interacts with the browser. In the graphical overlay that appears while the fragment is playing, the label is shown together with the

Table 1: Results for Experiment 1: User Incentive.

	Fragments		Feedback			Positive		Negative	
Visual Concept	$\sum$	%	$\sum$	%	*Avg*	$\sum$	%	$\sum$	%
Singer	117	23%	851	24%	7.3	782	26%	69	14%
Audience	87	17%	800	22%	9.2	697	23%	103	20%
Stage	82	16%	499	14%	6.1	462	15%	37	7%
Drummer	70	14%	560	16%	8.0	429	14%	131	26%
Guitar player	61	12%	335	9%	5.5	274	9%	61	12%
Close-up	35	7%	182	5%	5.2	174	6%	8	2%
Over the shoulder	20	4%	110	3%	5.5	102	3%	8	2%
Pinkpop logo	11	2%	93	3%	8.5	49	2%	44	9%
Pinkpop hat	10	2%	73	2%	7.3	48	2%	25	5%
Hands	9	2%	31	1%	3.4	23	1%	8	2%
Keyboard	8	2%	33	1%	4.1	17	1%	16	3%
Total	510	100%	3,567	100%	6.4	3,057	100%	510	100%

duration and thumbnails of the first, middle and last frame. With the thumbs-up and thumbs-down buttons the user indicates whether she agrees with the automatically suggested label or not. We refer to this feedback as a positive or negative user tag. If the user indicates that the label is incorrect, she may indicate to the system what the correct label for the video fragment should be, chosen from the predefined labels or a new suggestion from the user. The overlay disappears within a few seconds, but instantly after user-feedback, see Figure 3 for a typical feedback sequence.

3. EXPERIMENTAL SETUP

3.1 Field Study

The video search engine was available online from December 2009 to February 2010. In these three months almost ten thousand users, mainly from the Netherlands and Belgium, visited the site and used the system to watch concert videos with the in-video browser. Due to press releases by the different organizations involved, online media was the main source for traffic. A total number of 958 users provided feedback on the video fragment labels, which is a common participation rate for online communities [7]. From this user base, we classify 578 members as active users because they provided multiple tags. The average active user provided feedback on six different fragments, but two labels on different fragments is the most common. In this study we restrict ourselves to the positives and negative user tags. Additional feedback like suggestions for other labels or new start/end times are not taken into account for the current study. All user feedback was stored in a database together with the users IP addresses and user sessions.

3.2 Experiments

In contrast to standard retrieval evaluation tasks, where the user information need is captured in a standard model, our field study is done in a non-controlled environment, with real users and real-world user behavior. The effect is that little is known about the participating users, except for their IP-address and session-id. We conduct two experiments. In the first experiment: **User Incentive** we gather user tags and we evaluate whether the In-Video browser and the automatically labeled video fragments indeed encourage the users to provide feedback for free on a diverse set of concerts and fragments. The more challenging research question is whether users put enough effort in the tags so that they become reliable at an aggregated level. Therefore, the goal of the second experiment: **Quality vs Quantity**, is to verify the reliability of the user tags. This is performed by comparing the (aggregated) user tags with an expert-labeled ground truth.

4. RESULTS

4.1 Experiment 1: User Incentive

The question that we tried to answer with this experiment is whether a concept-based video search engine provides enough incentives for users to provide labels, without receiving payment or other compensation. We summarize the feedback statistics over the 11 visual concepts in Table 1. We received feedback on a total of 726 different fragments, but we focus our study on the 510 fragments that received at least two judgements. In total these fragments received 3,567 different tags distributed over 62 concerts. The feedback distribution shows that the four most common labels have received 75% of the user feedback. The feedback per fragment varies depending on the visual label from an average of 3.4 (Hands) to 9.2 (Audience).

There are multiple reasons why a fragment could have received more or less feedback: 1) presence of the visual concept varies per video, 2) detectors perform better for specific labels, 3) users could evaluate some labels better than others, and so on. In this non-controlled field study we could not collect all the data that is needed to analyze the dependencies of all those variables. It could be expected that users would be more easily motivated to provide negative feedback on incorrect suggested labels. We also investigated the relation between positive and negative judgements for each of the fragments (data not shown), from which we conclude that the user tags are independent of the quality of the automatically detected fragment label. Coming back to the question we started with, we conclude that within the context of our case study users had sufficient incentive to provide labels.

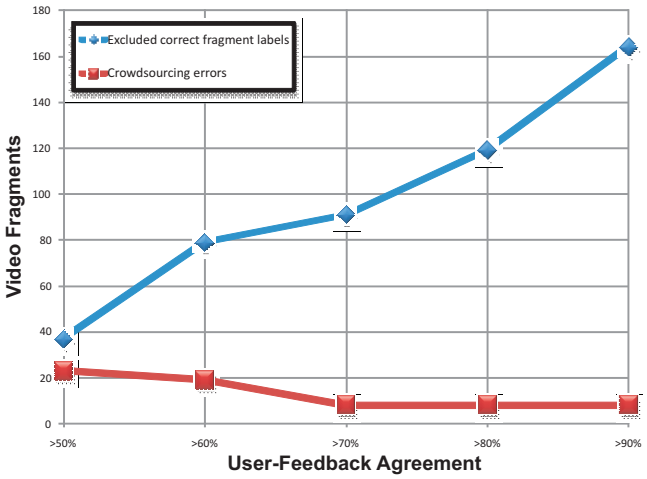

Figure 4: Results for Experiment 2: Quality vs Quantity. Simply relying on a majority vote of the crowd results in most correct fragments, albeit with 23 errors. We observe a best tradeoff between quality and quantity of crowdsourcing visual detectors for a user agreement of 67%.

4.2 Experiment 2: Quality vs Quantity

The question that we tried to answer with this experiment is whether the resulting labels are of sufficient quality compared to expert labels, when aggregated over multiple users. We have in total 510 fragments, where we now assume the expert label to be correct, and investigate for how many of them we would have obtained the same label when imposing a minimum agreement threshold on the crowdsourced labels. We plot the percentage of agreement among user-provided labels versus the number of video fragments in Figure 4. The ground truth shows that the quality of the suggested labels is high. As much as 85% of the automatically suggested labels correspond with the ground truth. If the simple evaluation principle of the majority is used, only 23 fragments have received tags that do not match with the ground truth, which in our case corresponds to a loss of 37 training samples. When we further increase the threshold for a positive or negative agreement the number of fragments receiving the wrong label is gradually reduced to 8 fragments only, but the number of excluded training samples increases rapidly. For a conservative user agreement of 80%, for example, 119 fragments are ignored. We observe that a threshold of 67% provides a well-chosen balance between the 8 errors and the 422 fragments that can be used as a correction mechanism, or as reliable training examples for a new round of detector learning.

5. CONCLUSION

The main research question of this paper was: can user tags from crowdsourcing be beneficial to a system that automatically predicts labels for video fragments. We developed a video search engine for a dedicated user community in the domain of concert video allowing for easy fragment-level crowdsourcing. The user-feedback mechanism of the In-Video browser made it possible to harvest positive and negative user judgements on automatically predicted video fragment labels.

For this case study two experiments are conducted. The first experiment showed that users provided enough feedback. Analysis of the collected data proved that users provided the feedback to the video-fragment labels without a preference for incorrect labels. The second experiment showed that 85% of the automatically suggested labels corresponds with the ground truth. We observe that an aggregation threshold of 67% provides a well-chosen balance between errors in the user judgements and the amount of reliable training examples remaining. If the threshold is enforced, the error rate in the training examples is less than 2%. Within the context of our case study, we conclude that crowdsourcing can be beneficial to enhance and improve automated video content analysis. How the new information can be exploited for incremental learning of visual detectors is an interesting question for future research.

6. ACKNOWLEDGMENTS

We thank our users for providing feedback. This research is supported by the projects: BSIK MultimediaN, FES COMMIT, and STW SEARCHER.

7. REFERENCES

[1] L. Ahn. Games with a purpose. *IEEE Computer*, 39(6):92–94, 2006.

[2] M. Ames and M. Naaman. Why we tag: Motivations for annotation in mobile and online media. In *Proc. CHI*, 2007.

[3] R. Gligorov, L. B. Baltussen, J. van Ossenbruggen, L. Aroyo, M. Brinkerink, J. Oomen, and A. van Ees. Towards integration of end-user tags with professional annotations. In *Proc. Web Science*, 2010.

[4] A. Kittur, E. H. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proc. CHI*, 2008.

[5] C. Marlow, M. Naaman, D. Boyd, and M. Davis. HT06, tagging paper, taxonomy, Flickr, academic article, to read. In *Proc. Hypertext*, 2006.

[6] P. Marsden. Crowdsourcing. *Contagious Magazine*, 18:24–28, 2009.

[7] J. Nielsen. Participation inequality: Encouraging more users to contribute, 2006. `http://www.useit.com/alertbox/participation_inequality.html`.

[8] D. A. Shamma, R. Shaw, P. L. Shafton, and Y. Liu. Watch what I watch: using community activity to understand content. In *Proc. MIR*, 2007.

[9] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. MIR*, 2006.

[10] C. G. M. Snoek, B. Freiburg, J. Oomen, and R. Ordelman. Crowdsourcing rock n' roll multimedia retrieval. In *Proc. ACM Multimedia*, 2010.

[11] C. G. M. Snoek and A. W. M. Smeulders. Visual-concept search solved? *IEEE Computer*, 43(6):76–78, 2010.

[12] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proc. EMNLP*, 2008.

[13] J. Surowiecki. *The wisdom of crowds: why the many are smarter than the few.* Random House, 2005.

[14] R. van Zwol, L. Garcia, G. Ramirez, B. Sigurbjornsson, and M. Labad. Video tag game. In *Proc. WWW*, 2008.