SocialZap: Catch-up on Interesting Television Fragments Discovered from Social Media

Svetlana Kordumova[1], Christoph Kofler[2], Dennis C. Koelma[1], Bouke Huurnink[3], Bauke Freiburg[4], Joris Kleinveld[5], Manuel van Rijn[5], Marco van Deursen[5], Martha Larson[2] and Cees G.M. Snoek[1]

[1]*Intelligent Systems Lab Amsterdam, University of Amsterdam, The Netherlands*
[2]*Multimedia Information Retrieval Lab, Delft University of Technology, The Netherlands*
[3]*Netherlands Institute for Sound and Vision, Amsterdam, The Netherlands*
[4]*Video Dock B.V., Amsterdam, The Netherlands*
[5]*Auxilium B.V., Amsterdam, The Netherlands*

ABSTRACT

In this paper we present *SocialZap*, a multimedia search engine that finds the most interesting fragments, *zap points*, in a television broadcast based on microblog posts and socially tagged photos. The main novelty of SocialZap is the fully-automatic transfer of the learned viewers interest from textual posts to the visual channel, without the need for any manual effort in the process. Once SocialZap finds the zap points, users can easily browse through a television broadcast and directly watch the interesting fragments. Thus, SocialZap adds social experience to watching television.

1. INTRODUCTION

Existing web services [1] allow viewers to watch missed television broadcast later on the web. However, a system that directly suggests the most interesting *fragments* to watch, based on social media, is, to the best of our knowledge, non existing. In our SocialZap demo, we analyze data from social media to suggest interesting concepts, zap points, in television broadcast. SocialZap uses textual information from Twitter posts related to a television broadcast of interest, which provide a rich source of information of what viewers find interesting, see Figure 1. The challenge we face is the temporal mismatch between the moment that the user tweets about a concept and the moment at which it appears in the television broadcast. The tweet-time can radically differ from the appearance-time as viewers either anticipate appearances or continue to tweet about topics that have previously appeared. To transfer the found concepts of interest from the textual to the visual modality, we use state of the art approach [8] which learns concept detectors from social media. We collect Flickr images tagged with the concept names, and we select the most reliable ones as training data for learning. Thus, the proposed system relies totaly on social media data, without any manual annotations in the process. The SocialZap pipeline is illustrated in Figure 2, we describe it in more detail in the following section.

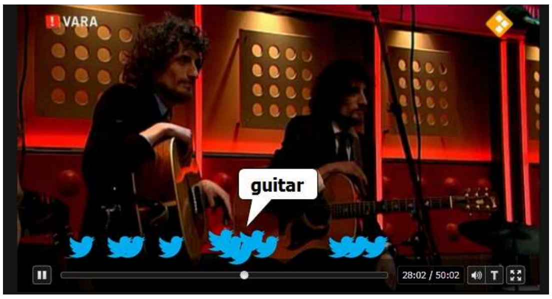


Figure 1: Video zap points, corresponding to the time points within a television broadcast at which the visual appearance of concepts designated by common terms extracted from tweets, have been detected.

2. SYSTEM OVERVIEW

In order to automatically provide zap points to users, SocialZap obtains *interesting* (visual) topics about the television broadcast and then learns visual concept models to recognize those concepts in an unseen video. SocialZap exploits social information from (1) Twitter posts related to a television broadcast of interest to find the most common terms as concepts of interest and (2) tagged Flickr images to learn visual concept detectors. SocialZap is composed of different modules, which communicate with each other through a common API in a central server, see Figure 3. We describe the separate modules in the following sections.

2.1 Textual Concepts and Social Media

Assuming that viewers write posts about what they find interesting, we exploit social information about the television broadcast gathered from Twitter with the Tweet Analyzer module of SocialZap. The posts are collected by issuing queries with automatically constructed *hashtags* from the broadcast show name, as *#dwdd* shown in Figure 2. The tweets that are posted during the time of the broadcast are considered to be associated with the broadcast. Additionally we exploit closed captions associated with the television broadcast with the Caption Analyzer module. Closed captions are beneficial to collect visually depicted entities if

ICMR'14, April 01–04 2014, Glasgow, United Kingdom.

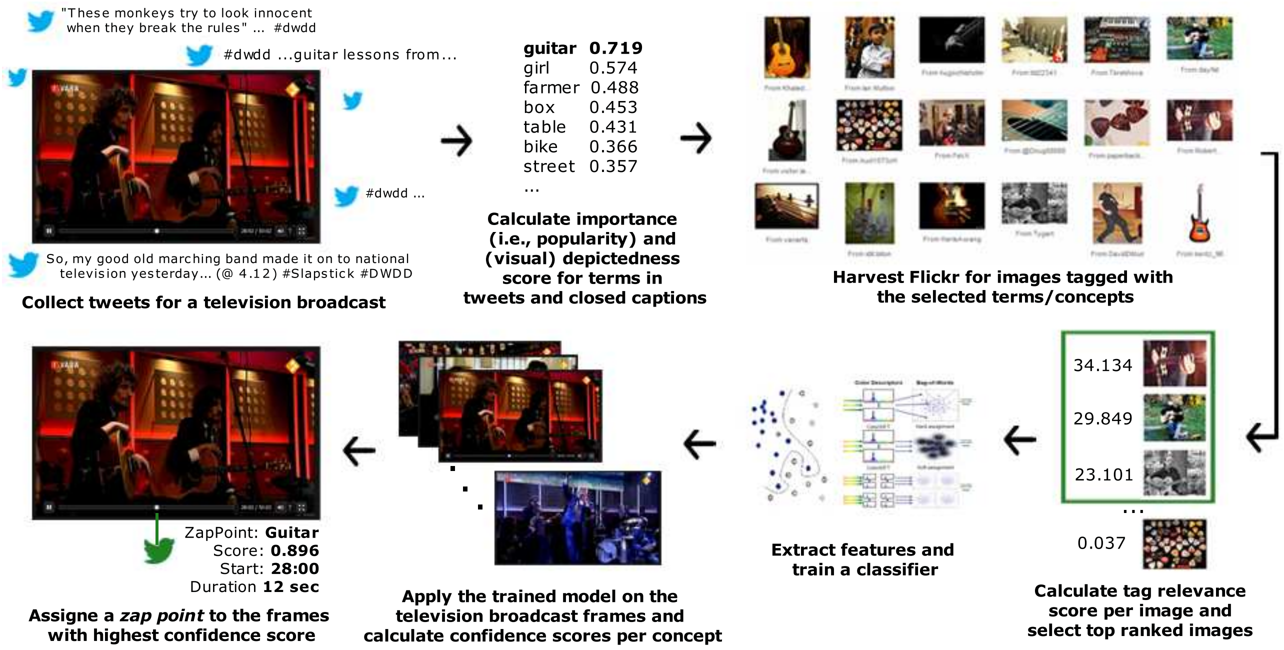

"These monkeys try to look innocent when they break the rules" ... #dwdd
#dwdd ...guitar lessons from...
#dwdd ...
So, my good old marching band made it on to national television yesterday... (@ 4.12) #Slapstick #DWDD
Collect tweets for a television broadcast
guitar 0.719
girl 0.574
farmer 0.488
box 0.453
table 0.431
bike 0.366
street 0.357
...
Calculate importance (i.e., popularity) and (visual) depictedness score for terms in tweets and closed captions
Harvest Flickr for images tagged with the selected terms/concepts
34.134
29.849
23.101
...
0.037
Calculate tag relevance score per image and select top ranked images
Extract features and train a classifier
Apply the trained model on the television broadcast frames and calculate confidence scores per concept
ZapPoint: Guitar
Score: 0.896
Start: 28:00
Duration 12 sec
Assigne a zap point to the frames with highest confidence score

Figure 2: SocialZap pipeline. Tweets for a particular television broadcast are gathered and visual detectors for the most common terms in the tweets are then learned to find the visual fragments of those concepts as zap points.

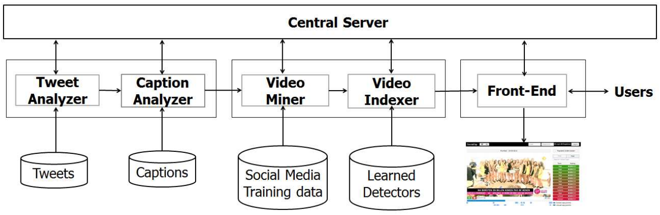


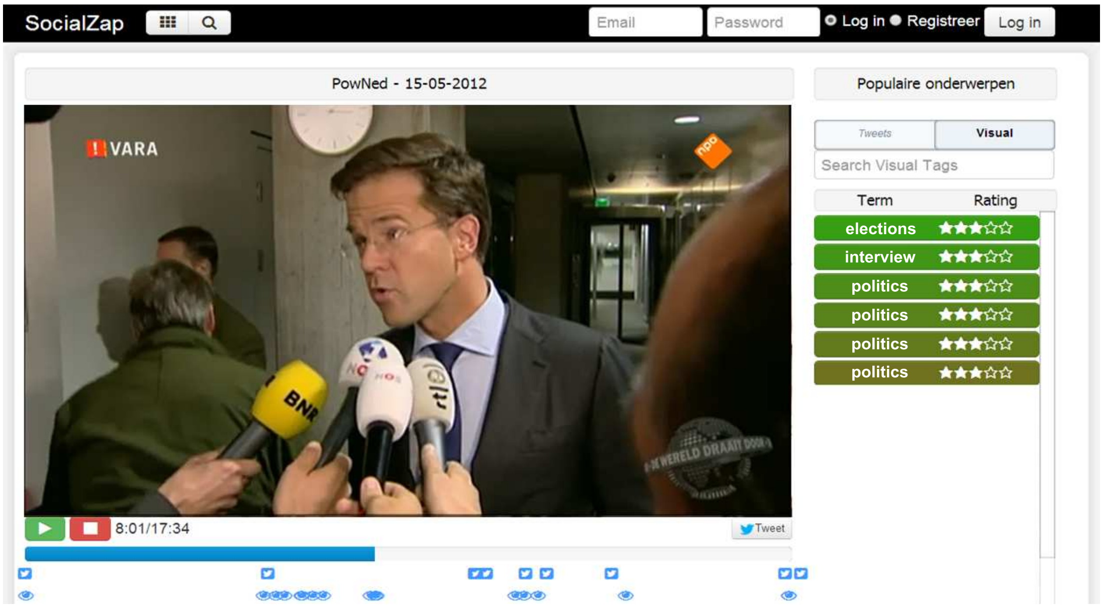
Figure 3: SocialZap is composed of different modules, which communicate with each other through a common API in a central server. The Tweet and the Caption Analyzer are responsible for finding common terms. The Video Miner learns video concept detectors for the depicted terms, using training data only from Social Media. The Video Indexer finds the zap points by applying the learned concept detectors on television broadcasts.

there are not enough found in the tweets, and also to denoise tweet-terms by assuring that they also occur at least once in the captions. After obtaining tweets and closed captions, we translate them to English and perform stopword removal. This preprocessing step is necessary to build the basis for later retrieving more accurate training data for selected visual concepts. Each of the terms contained in tweets and closed captions gets assigned a reranking score in the interval $[0, 1]$, which represents both the *importance* (i.e., popularity) and *(visual) depictedness* of these terms with respect to the harvested tweets and the (audiovisual) content of the corresponding broadcast. The *importance* of a term is obtained by building language models [11, 2] for each television broadcast. Terms that are specific to a television broadcast and at the same time mentioned relatively frequently in Twitter posts receive a higher relevance score. In order to quantify the *depictedness* of a term, we apply the generic approach from [4]. The overall score of a term is modelled by combining importance- and depictedness-scores in a linear fashion.

2.2 Visual Concepts and Social Media

To transfer viewers interest from the textual modality to visual concepts, we need training data to learn visual concept detectors. In order to detect any possible video concept, a promising line of research is to automatically collect training examples from the web, where many socially tagged videos and images exist [5, 9]. In a recent study, we have shown that tagged images are a better training source than tagged videos for learning video concept detectors [3]. Therefore we use Flickr as source to collect tagged images. Due to variation between users regarding the use of social tags, simply treating all tagged images as positive examples may be problematic [5]. In order to preserve only good examples, we first calculate tag relevance scores per image. We use the multi-feature variant of the neighbor voting algorithm [5]. The images are then ranked in terms of their estimated tag relevance scores. Then, we preserve a top ranked proportion of the examples [3], as shown in Figure 2. As negative examples, we employ *negative bootstrap* [6], which finds relevant negative examples in an iterative manner. We follow a state-of-the-art bag of visual codes pipeline for feature extraction [10], and learn concept detectors using one-vs-all SVM [7]. We call this module of the framework Video Miner. The described approach implemented in SocialZap showed best performance on the TRECVID 2013 semantic indexing with no annotation task [8], with best score for 31 out of 38 concepts.

Once we obtain the learned concept detectors, we apply them on the broadcasted video frames. For each frame we get confidence scores on how likely the concept appears in the frame. The frames with highest scores are suggested as



PowNed - 15-05-2012

Populaire onderwerpen

Term	Rating
elections	★★★☆☆
interview	★★★☆☆
politics	★★★☆☆
politics	★★★☆☆
politics	★★★☆☆
politics	★★★☆☆

Figure 4: User Interface of SocialZap. Once the user selects a show, all found zap points are listed on the right side of the video player, ranked by importance. The user can select any zap point of interest. Once selected the interface starts playing the video at the frame where the zap point occurs.

zap points of the television broadcast. This module is the Video Indexer.

2.3 User Interface

We created a simple user interface that allows users to *catch-up* on the interesting fragments of a broadcast, see Figure 4. The interface provides an option to select a specific television show. Once selected, SocialZap lists the most common terms found in the social media posts with respect to the selected broadcast. When the user chooses one of the terms, the video player directly plays the video segment where the significance score for that concept is highest. If some term occurs more then once in the video, all zap points for that term will be listed, as in the case with the term *politics* shown in Figure 4. As it can be seen from the figure, the concept detector of *elections* sets the zap point correctly at a relevant frame where this concept appears. However since concept detection is a challenging task and does not always give precise results, we found that timing of tweet and caption terms can help for finding correct zap points.

3. CONCLUSION

In this paper we present the SocialZap system, which finds interesting fragments in television broadcasts based on social media. With SocialZap the viewer's interest is depicted from Twitter and transferred into the visual channel as zap points. The system provides an easy way to browse through the zap points of the television broadcast, adding a new social dimension to the television experience.

Acknowledgments. This research is supported by the Dutch national program COMMIT, project SEALINCMedia. We thank the students from TU Delft for creating the user interface of Social Zap.

4. REFERENCES

[1] Uitzendinggemist, online video portal of the Dutch public broadcasters, *http://www.uitzendinggemist.nl/*.

[2] S. Cronen-Townsend, Y. Zhou, and W. B. Croft. Predicting query performance. In *ACM SIGIR*, 2002.

[3] S. Kordumova, X. Li, and C. G. M. Snoek. Evaluating sources and strategies for learning video concepts from social media. In *CBMI*, 2013.

[4] M. Larson, C. Kofler, and A. Hanjalic. Reading between the tags to predict real-world size-class for visually depicted objects in images. In *ACM MM*, 2011.

[5] X. Li, C. G. M. Snoek, and M. Worring. Unsupervised multi-feature tag relevance learning for social image retrieval. In *CIVR*, 2010.

[6] X. Li, C. G. M. Snoek, M. Worring, D. C. Koelma, and A. W. M. Smeulders. Bootstrapping visual categorization with relevant negatives. In *TMM*, 2013.

[7] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machines is efficient. In *CVPR*, 2008.

[8] C. G. M. Snoek, K. E. A. van de Sande, D. Fontijne, A. Habibian, M. Jain, S. Kordumova, Z. Li, M. Mazloom, S.-L. Pintea, R. Tao, D. C. Koelma, and A. W. M. Smeulders. MediaMill at TRECVID 2013: Searching concepts, objects, instances and events in video. In *TRECVID*, 2013.

[9] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. Identifying relevant frames in weakly labeled videos for training concept detectors. In *CIVR*, 2008.

[10] K. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *PAMI*, 32(9):1582–1596, 2010.

[11] C. Zhai. Statistical language models for information retrieval a critical review. *FnTIR*, 2(3):137–213, 2008.