Künstl Intell (2014) 28:333–336
DOI 10.1007/s13218-014-0328-x

DOCTORAL AND POSTDOCTORAL DISSERTATIONS

Visual Learning of Semantic Concepts in Social Multimedia

Damian Borth

Published online: 24 September 2014

1 Introduction

Currently, traditional media is experiencing a major shift towards social media. At the same time, interaction via social media is to an increasing degree enriched with images and videos, as seen during the Arab Spring in the Middle East in 2012 or the Boston Marathon Bombings on April 15, 2013. This combination gives rise to a new type of content, which is being called social multimedia.

One key trigger for this trend is the capability to upload and distribute images and videos over the Internet minutes after such incidents happen on video-sharing platforms like YouTube or Vimeo. This is possible due to the availability of broadband Internet, the low price of storage, and the omnipresence of camera-equipped mobile devices. It allows people to record, publish, and share digital images and videos without notable effort. This ubiquity of visual content conveys much about our thinking and feeling, in that it reflects our personal lives and ourselves as a society.

Unfortunately, this content is of little use if it is not accessible to users, e.g., by allowing users to retrieve videos by keyword-based search. However, keyword-based search requires each individual video to be annotated with a set of keywords describing its content. Given the vast amount of video content being created nowadays (YouTube, for example, stores about 100 hours of video content every minute) this poses an impossible task for human annotators.

Worse, as naturally as humans can perceive their surroundings visually, this undertaking is quite challenging for machines. This lack of correspondence between the low-level features that machines can extract from videos (i.e., the raw pixel values) and the high-level conceptual interpretation a human associates with perceived visual content is referred to as the semantic gap [6].

In recent years, great effort has been spent on content-based methods directly analyzing the video stream to bridge this gap. Following this line of research, the thesis focuses on concept detection [7], the task to detect semantic concepts in visual content. Given an input video clip, concept detection systems use statistical learning to infer the presence of a target concept by calculating its probability of appearance from low-level features extracted from the content. For this purpose, the set of all concepts—or concept vocabulary—should cover a broad spectrum of entities, such as objects (“chair”, “telephone”), scene types (“cityscape”, “desert”), and activities (“interview”, “people singing”), requiring concept detection systems to provide detectors for hundreds or even thousands of target concepts.

This, however, is considered as a major challenge in concept detection, as it demands labeled training samples for supervised machine learning—the underlying technology of current systems [7].

Such ground-truth training samples are usually acquired manually, i.e., a human annotator labels videos for whether the concept occurs. This time-consuming and cost-intensive effort creates a scalability problem, leading to small-scale, fixed concept vocabularies being useful in research setups, but making it impossible to satisfy the changing demands of users’ information needs. This leads to state-of-the-art systems still focusing on generic concepts such as “quadruped” or “hand” instead of providing detectors for concepts of interest, e.g., sports events such as “Olympics 2012”, incidents such as the “Costa

D. Borth (✉)
International Computer Science Institute & UC Berkeley,
1947 Center Street, Ste. 600, Berkeley, CA 94704, USA
e-mail: borth@icsi.berkeley.edu

Fig. 1 An illustration of the proposed concept detection approach. An unknown video clip is analyzed to automatically identify different types of labels: trending topics (*pink*), concepts (*blue*), and sentiment (*green*)

Concordia" accident, or product releases such as the new "iPhone". Finally, while there are approaches that can infer affect in visual content, no methods have yet been described in the literature for sentiment prediction from visual content. However, this kind of automatic assessment would lead to more comprehensive descriptions of social multimedia, where people express their opinions and sentiments on a regular base.

The thesis [2] presents strategies to address the above outlined challenge by proposing a novel combination between visual learning of semantic concepts and social media analysis: first, social media streams are mined for trending topics to synchronize concept detection with real world events matching users' information needs. Second web video from platforms such as YouTube is exploited as an alternative training source for concept detection. Youtube's user-generated tags are used as positive labels for supervised machine learning. Third, concept detection is extended by a large-scale visual sentiment ontology (VSO). The resulting SentiBank detectors are constructed from the analysis of emotions expressed on YouTube and Flickr.

The proposed framework is illustrated in Fig. 1: given an unknown video, the framework can be used to analyze visual content on different levels such as semantic concept (blue), trending topic (pink), and sentiment (green). In sum, the thesis allows concept detection to be more dynamic with respect to concept vocabulary and detector training and provides an understanding of how adjectives influence semantic concept when they are combined as concept pairs.

2 Dynamic Vocabularies from Trending Topics

The first contribution of the thesis is its novel approach towards forming dynamic vocabularies for concept detection. The key idea is to expand concept vocabularies with trending topics that are mined automatically from media like Google, Wikipedia, and Twitter [4]. To achieve this, topics from different media channels are clustered and aggregated to form daily trending topics (see Fig. 2).

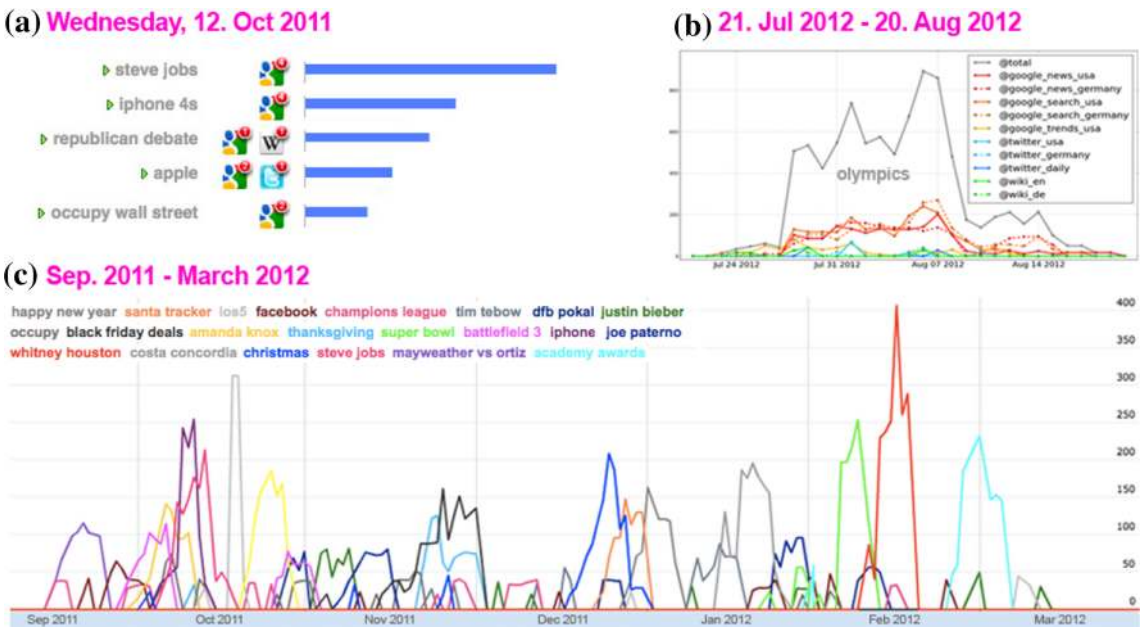

Fig. 2 **a** For each day, the top trending topics are identified by aggregating feeds from Google, Wikipedia, and Twitter, and trend scores are computed. **b** Individual trend scores for the topic "Olympics 2012" during summer 2012. **c** The trend scores for the 23 most prominent trends for September 2011–March 2012, plotted over time

Following, the thesis presents the first comprehensive study of various trending topic characteristics across three major online and social media streams, covering thousands of trending topics during an observation period of an entire year. Results from this study show that a typical trending topic "lives" for up to 14 days, with an average of 5 days. Surprisingly, the analysis indicates that Wikipedia as a media channel is as quick as Twitter when it comes to the first appearance of a trending topic.

An important condition for constructing concept vocabularies dynamically is the capability to predict the most popular trending topics for detector training. This is done by forecasting the life-cycle of trending topics at the very moment they emerge. The presented fully automated approach is based on a nearest-neighbor forecasting technique, exploiting the assumption that semantically similar topics exhibit similar behavior [1]. In experimental results, it is shown that this approach is able to forecast the progression of trending topics more accurately than state-of-the-art auto-regression moving-average methods.

Once identified, the trending topics can be either mapped to a static concept vocabulary, trained as a single detector on-the-fly, or used to expand an existing static vocabulary. These three strategies for establishing dynamic vocabularies are evaluated on 6,800 YouTube videos and the top 23 target topics from the dataset. The results show that direct visual classification of trends (by a "live" learning from trending topic videos) outperforms inference from static vocabularies, and can further be improved by a combination of the first and second strategy.

3 Web Video and Active Relevance Filtering

One major challenge in concept detection based on web video is that of how to retrieve proper visual training content from web platforms like YouTube. Similar to a textual web search, where a user has to define a set of keywords to formulate a search query, a concept detection system must define a proper set of keywords for API query construction e.g. for a concept like "car race", the videos retrieved should not include remote-controlled cars or interviews with racecar drivers. The thesis presents an approach that offers an automatic concept-to-query mapping for training data acquisition from YouTube, where queries are automatically constructed by keyword selection and category assignment. Results demonstrate that the proposed method allows the system to automatically retrieve training videos, that are as relevant as those retrieved manually by humans.

Despite these improvements, and due to the coarseness of web video tags, these videos only serve as weak indicators of concept presence. Such weakly labeled web video contains lots of non-relevant content. So far, there are two general strategies for overcoming this problem: (1) manual refinement supported by active learning sample selection, and (2) automatic refinement using relevance filtering [8]. The thesis presents an approach combining these two strategies in an interleaved setup (Fig. 3): manually refined samples are directly used to improve relevance filtering, which in turn provides a basis for the next active learning sample selection. Results demonstrate that the proposed combination—called active relevance filtering—outperforms both, a purely automatic filtering and a manual filtering based on active learning.

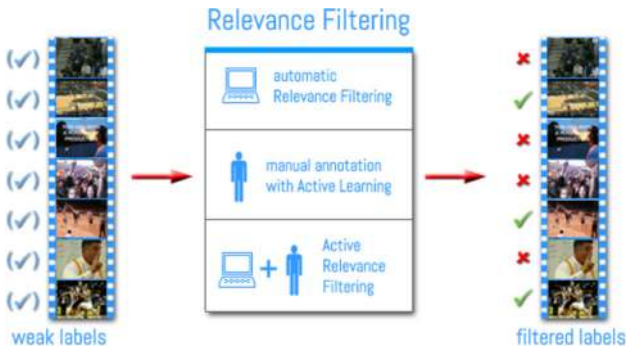

Fig. 3 Overview of difference label refinement strategies for weakly labeled web video

4 Adjective Noun Pairs for Visual Sentiment

The third contribution of the thesis is to tackle the challenge of sentiment analysis based on visual content as illustrated in Fig. 4.

This is accomplished by introducing a large-scale ontology of 3,000 adjective noun pairs (ANPs) [3]. This ontology is based on psychological theory and the proposed construction method is fully data-driven, i.e., it automatically mines online sources such as Flickr and YouTube for sentiment words, which serve as the building elements for

Fig. 4 Barack Obamas reelection tweet (*left*) and a tweet capturing the destruction caused by Hurricane Sandy (*right*). Both tweets convey their main information, including the sentiment, visually

Fig. 5 Adjective Noun Pair examples illustrating how adjectives can change the sentiment a phrase conveys, and how this can be used to differentiate the visual space corresponding to a noun such as "dog"

ANPs discovery. Further, it presents SentiBank, a novel mid-level representation framework, which is built on the ontology and encodes visual presence of 1,200 ANPs. This bank of concept detectors is able to differentiate between visual concepts such as "cute dog" and "dangerous dog" (Fig. 5) and therefore provides an unique understanding of visual sentiment.

In experiments on sentiment analysis with real-world Twitter data covering 2,000 photo tweets, the proposed mid-level representation demonstrates an improved prediction accuracy of 13 % (absolute gain) in a joint visual-text setup over a state-of-the-art text only methods.

It has also been demonstrated that this approach is able to outperform state-of-the-art porn detection baselines [5] on real-world pornographic and child sexual abuse (CSA) content. Additionally, the compilation of detected ANPs allows to explain detection results to law-enforcement, which in this domain is an important system requirement.

To summarize, the presented visual sentiment analysis effort—being the first of its kind—has created a large, publicly available resource consisting of a concept ontology, a detector library, and a training/testing benchmark for visual sentiment analysis.[1]

[1] http://visual-sentiment-ontology.appspot.com.

References

1. Althoff T, Borth D, Hees J, Dengel A (2013) Analysis and forecasting of trending topics in online media streams. In: Proceedings of ACM multimedia
2. Borth D (2014) Visual learning of socio-video semantics. Ph.D. thesis, University of Kaiserslautern, Germany
3. Borth D, Ji R, Chen T, Breuel T, Chang SF (2013) Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: Proceedings of ACM multimedia
4. Borth D, Ulges A, Breuel TM (2012) Dynamic vocabularies for web-based concept detection by trend discovery. In: Proceedings of ACM multimedia
5. Schulze C, Henter D, Borth D, Dengel A (2014) Automatic detection of child pornography by multi-modal feature fusion for law enforcement support. In: Proceedings of ACM multimedia retrieval (ICMR)
6. Smeulders A, Worring M, Santini S, Jain AGR (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380
7. Snoek CGM, Worring M (2008) Concept-based video retrieval. Found Trends Inf Retr 2(4):215–322
8. Ulges A, Borth D, Breuel T (2010) Visual concept learning from weakly labeled web videos. In: Video search and mining, Springer

Damian Borth received his doctoral degree from the University of Kaiserslautern, where he was a member of the Multimedia Analysis and Data Mining Group at the German Research Center for Artificial Intelligence (DFKI). He was a visiting researcher in the Digital Video and Multimedia lab at Columbia University in 2012 and is currently at UC Berkeley and the International Computer Science Institute (ICSI) in Berkeley, USA. His research interests includes visual learning, multimedia retrieval and social media analysis.