Towards Fusion of Collective Knowledge and Audio-Visual Content Features for Annotating Broadcast Video

Fréderic Godin
Multimedia Lab
Ghent University - iMinds
Ghent, Belgium
frederic.godin@ugent.be

Wesley De Neve
Multimedia Lab
Ghent Univ. - iMinds & KAIST
Ghent, Belgium
wesley.deneve@ugent.be

Rik Van de Walle
Multimedia Lab
Ghent University - iMinds
Ghent, Belgium
rik.vandewalle@ugent.be

ABSTRACT

Broadcasters produce vast collections of video content. However, the lack of fine-grained annotations makes it difficult to retrieve video fragments of interest from these vast collections. Indeed, manual annotation of video content is labour-intensive and time-consuming. Moreover, the applicability of algorithms for automatic annotation of video content is limited, given that too many prerequisites need to be fulfilled and that a lot of concepts are unidentifiable. At the same time, people are using social media to share their thoughts about the content they view on television. Therefore, in this Ph.D. research, we plan to investigate novel machine learning-based approaches towards the task of fine-grained annotation of broadcast video content, fusing the collective knowledge present in social media with the output of audio-visual content analysis algorithms.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.5.4 [**Pattern Recognition**]: Applications

General Terms

Algorithms, Design, Human factors

Keywords

Annotation; broadcast video; collective knowledge; content analysis; multi-modal fusion; signal processing; social media

1. INTRODUCTION

To facilitate keyword-based search in large collections of video content, retrieval algorithms make use of textual descriptions. These textual descriptions can be added manually or automatically. Manual annotation of video content is a time-consuming and labour-intensive task. Therefore, researchers have developed audio-visual content analysis algorithms that are able to annotate video sequences automatically. To that end, mappings have been defined between low-level audio-visual content features and high-level semantic concepts. Despite promising results, two major problems can be identified:

- **Generalizability:** Visual concept detectors are often a combination of carefully engineered features and heuristic rules that are tailored on the training data. Therefore, as Yang and Hauptmann [10] prove, these approaches are not generalizable to similar content from other sources (or to other domains), and depend highly on the content prerequisites that need be fulfilled.

- **Concepts with limited audio-visual (AV) characteristics:** To be able to correctly detect a concept, that concept needs to exhibit discriminative AV characteristics. Otherwise detection becomes cumbersome. E.g., in soccer video sequences, when detecting yellow cards using AV features, a significant number of false positives are generated because of the lack of highly discriminative AV features [3].

In recent years, social media generated a substantial amount of collective knowledge [4], providing context for content. Therefore, we hypothesize that the solution to the problem of automatic annotation of video content lies in combining algorithms for collective knowledge analysis and algorithms for audio-visual content analysis. Indeed, the content prerequisites can be fulfilled by taking advantage of the collective knowledge available in social media. Moreover, previously undetectable concepts can be annotated, again using the collective knowledge available in social media. As such, the problem statement changes from how to detect semantic concepts to how to align content and context, namely multimedia objects and collective knowledge.

In this Ph.D. research, we aim at answering the following research question: how to make use of collective knowledge for the purpose of automatically annotating broadcast video at a fine-grained level, given that collective knowledge allows taking advantage of a rich concept vocabulary on the one hand, and that collective knowledge is often unstructured, multilingual, subjective, and noisy on the other hand?

This paper is structured as follows. In Section 2, we discuss related work in the field of fusion. We also briefly discuss the novelty of this Ph.D. research with respect to the related work. Next, in Section 3, we propose a novel method for fusing collective knowledge with AV features, with the overall aim of annotating broadcast video. In Section 4, we present initial experimental results. Finally, in Section 5, we provide conclusions and directions for future research.

ICMR'13, April 16–19, 2013, Dallas, TX, USA.

2. RELATED WORK

Our initial experiments make use of Twitter (see Section 4). Therefore, our review of related work consists of a discussion of algorithms to analyze video content and unstructured textual data, and of techniques to fuse both. In Section 2.1, we discuss multi-modal fusion. In Section 2.2, we outline existing approaches for fusing video content and Twitter data. Finally, in Section 2.3, we detail the novelty of the proposed Ph.D. research.

2.1 Multi-modal fusion

In multi-modal fusion, complementary modalities are fused to increase the accuracy of the overall decision-making process [1]. Given that different streams of data can have different processing times, formats, or costs, most researchers opt for a late fusion approach that is layered in nature, first analysing the data streams independently and subsequently fusing the results obtained for the different streams.

Xie *et al.* [8] use a Hierarchical Hidden Markov Model (HHMM) in order to fuse audio and video streams. In addition, probabilistic Latent Semantic Analysis (pLSA) is used to cluster text transcript by topic. Both algorithms yield a mid-level semantic representation of their corresponding stream. Next, the mid-level representations are clustered, using a dynamic mixture model to group video fragments and their descriptions by topic. Similar techniques were applied in [9], with the aim of aligning webcast text and videos of sports games. In particular, pLSA is used for text-based clustering of events. Next, Conditional Random Fields (CRFs) are used to obtain event clusters that are aligned in time with segmented video sequences. The authors stress the importance of the use of sequential models because a lot of information is contained in previous and following states/topics/fragments.

2.2 Fusing Twitter and video streams

Given the scale and real-time nature of Twitter, event detection in Twitter is a highly popular research topic. A Twitter stream can be seen as a free stream of collective knowledge about events that happen in the world. When a broadcast video stream is available, microposts about certain fragments can be used to browse through the video and retrieve certain fragments.

Shamma and Churchill [5] use tweets to annotate broadcast videos of the inauguration of President Obama and the presidential debates of 2008. They apply the Term Frequency - Inverse Document Frequency (TF-IDF) principle in order to extract keywords that facilitate video browsing. In addition, they calculate Importance and Chatness scores to identify the most important events and the most discussed events in the video sequence, respectively. A similar idea is applied in [6], harnessing the *Wisdom of the Crowd* to detect the most important moments in live video streams for archiving purposes. The authors take into account the time between keyword occurrences, the cosine similarity between tweets, and the follow-followee relationships between Twitter users.

Highlight extraction from video sequences is also a very popular research topic. This holds particularly true for sports games. Lanagan and Smeaton [3] demonstrated that Twitter-driven detection of significant events in soccer and rugby games, combined with shot boundary detection, performs as good as event detection solely using audiovisual content analysis. For indexing and retrieval purposes, the authors make use of the bag-of-words model. This is done in a rough manner, using a timespan of at least one minute. To allow for more personalized and specialized event detection, the authors of [7] make use of spike detection and Support Vector Machines to detect and classify five types of soccer events and to discriminate between tweets from different teams, thus facilitating personalized, team-based video summarization.

2.3 Novel contributions

The big drawback of all previously described methods is that none of these methods uses video information to index the video content. All of these methods focus on extracting useful information of Twitter streams without taking into account the information contained within the video stream itself, despite the fact that researchers focused for more than a decade on video content analysis. The content and context are analysed separately. The rise of social media, however, makes it possible to take advantage of new information (context) about this content. Exploiting this content-context relationship gives several advantages, such as confidence-based algorithm selection using the context and exploiting the correlation between multiple streams.

To conclude, the novelty of the proposed Ph.D. research is as follows:

- Fusion and alignment of social media, especially Twitter, and video content, using multiple sources of information (see Section 3.2).
- Exploitation of the content-context relationship for algorithm selection using path planning (see Section 3.3).
- Provisioning of fine-grained non-expert annotations instead of coarse minute-by-minute descriptions.

3. PROPOSED METHOD

The goal of this Ph.D. research is to develop a collective knowledge-based framework for annotating broadcast video. The starting point is the Twitter data that are related to the broadcast video, and where the Twitter data form an unstructured, multilingual, noisy, and subjective but near-synchronous description of the broadcast video. Other sources of collective knowledge can be chosen too, or can be queried based on extracted information from the Twitter data. Think for instance about the retrieval of similar video clips from YouTube or the retrieval of similar images from Flickr or Facebook.

A schematic overview of the proposed framework is given in Figure 1. At the input side, we find the broadcast video (1). Based on what it knows about the broadcast video (e.g., the title of a show), the selection component will issue a number of queries (2) for initial information about the broadcast video (3). In our case, we will query Twitter in order to get a first timestamped description of the video. Next, an algorithm is selected that will fuse both the collective knowledge and the broadcast video, for instance making use of text processing techniques and computer vision algorithms to deduce new information (4). Based on this new information (5), we can again query sources of collective knowledge (2). An iterative process will take place until the selection component decides that a state of convergence has been reached. The output is an annotated broadcast video

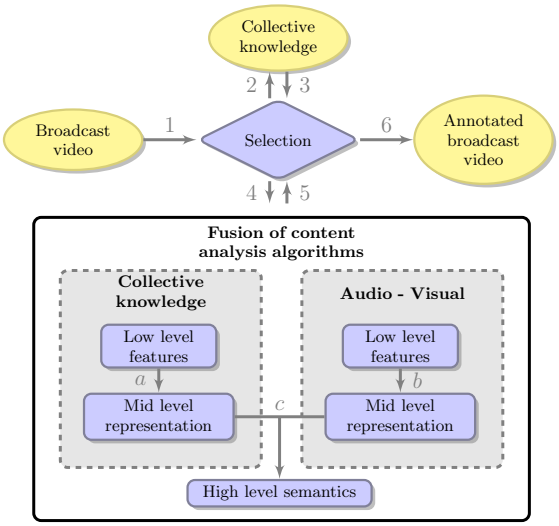

Figure 1: Schematic overview of the annotation framework.

containing fragment-by-fragment descriptions that are able to facilitate more effective video retrieval (6).

3.1 Collective knowledge

Collective knowledge can be defined as unstructured knowledge, generated by users of social media who often do not have expert knowledge on the topic. This knowledge can take the form of text, images, and video sequences. Although the freely-available collective knowledge empowers us with massive amounts of new information, we can also identify a number of disadvantages. When collective knowledge takes the form of text (e.g., tags or micro posts), natural language or multiple languages can be used. In addition, the information may be opinionated, and even incorrect.

Therefore, we can identify the following tree major research challenges regarding the use of collective knowledge: (1) how to structure and filter the data?, (2) how to deal with noise?, and (3) how to deal with sentiment?

As mentioned before, our starting point will be the Twitter stream associated with the broadcast video. In other words, our starting point will be the observations, remarks, and opinions of the viewers of the broadcast video. It can be expected that these micro posts will contain a significant number of named entities, like persons, objects, organizations, and locations. The verification of the presence of these named entities in the broadcast video is one possibility to deal with the aforementioned research challenges.

3.2 Fusion of content analysis algorithms

One of the most interesting ways to verify the information in collective knowledge is to exploit the correlation between the collective knowledge and the broadcast video. If we have high confidence in the alignment and correctness of a part of the data, we can accept the non-verifiable data as correct too.

Currently, no fusion approach exists that aims at fusing a broadcast video and a corresponding Twitter stream, despite the fact that previous work has already shown that the use of external text sources in a fusion approach to annotate sports video content can yield outstanding results [9].

In this Ph.D. research, we will apply a hybrid layered fusion approach. Different data streams will first be handled separately because of their different representation formats to yield a more uniform mid-level semantic representation (a and b). When two data streams are asynchronous, it should be clear that the fusion of these data streams will not be obvious. Therefore, at the highest level, the mid-level representations of several data streams can be fused using a sequential machine learning approach (c). To that end, Hidden Markov Models or Conditional Random Fields can be used, as well as a clustering approach based on Dynamic Mixture Models [8]. Collective knowledge fragments (events) and video fragments (shots) will be clustered based on the mid-level representations.

The mid-level representations are deduced from both the collective knowledge and the video sequence. There is no strict separation between them. For example, in order to be able to identify people in a video sequence by means of computer vision algorithms, we can infer an initial database of people from the Twitter data. Existing computer vision algorithms will be tailored to work with collective knowledge and new techniques will be developed to filter collective knowledge. To process vast streams of textual data, we can make use of topic models such as pLSA [8, 9] . The highest ranked keywords within a topic can be used to describe the content. Because collective knowledge is often subjective and opinionated, there is a need for sentiment analysis. This type of analysis has shown to be effective in selecting highlights during live broadcasts (see Section 4).

3.3 Selection component

The selection component is responsible for guiding the annotation process. It is an iterative process, where, based on the available information, new information sources can be queried. These new sources of information may take the form of external sources of collective knowledge. However, these new sources of information may also take the form of internal sources, namely the output of algorithms.

This process can be described as the use of context to decide how to annotate best the content. Currently, researchers make use of if-else constructs to do this [1], an approach that is not scalable. It would be better to let the algorithms learn for themselves how to obtain good annotations by means of machine learning-based techniques, such as reinforcement learning or online learning. In the case of reinforcement learning, confidence measurements about the new annotations could be used to determine the best path. That way, the selection component will learn the best path itself, rather than using of a set of human-defined rules.

4. EXPERIMENTAL RESULTS

As a first experiment, we extracted highlights from soccer video sequences of the English Premier League 2011-2012 in real time, using the framework shown in Figure 2. We made use of a two-step algorithm to detect six important events (goal, penalty, foul, substitution, end of first half, end of second half) in the Twitter streams associated with the soccer games. First, a sliding window approach is applied in order to detect spikes (4). Next, tweets are classified within a spike using a Support Vector Machine (5). Because highlights are typically highly emotional moments, we made use

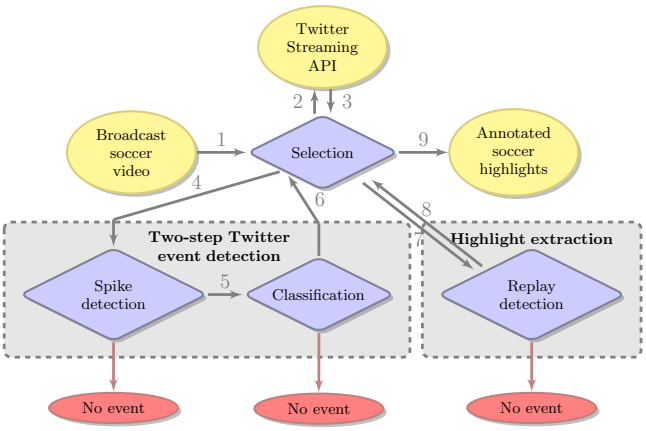

Figure 2: Instantiation of the proposed framework: highlight detection and extraction in soccer videos using Twitter

of features that resemble sentiment, such as the use of multiple exclamation marks or the capitalization of keywords and player names. The presence of keywords belonging to an event class were also added to be able to discriminate between events. Different from previous approaches [3, 2, 7], we made use of the video content to verify the correctness of the detected events. After the detection of an event on Twitter (6), the selection component decides to seek the corresponding fragment in the video sequence (7). Currently, replay detection is applied within a certain time span around the point of detection making use of logos. When found (8), an annotated replay fragment is delivered (9).

We evaluated our algorithm on detecting goals in 12 soccer games that contained 34 goals. Our algorithm obtained a precision and recall of 93.75% and 88.26%, respectively. This shows that video content analysis can benefit from external sources of collective knowledge. The time and type of the detected event allows us to define content prerequisites for applying video content analysis methods, making it possible to significantly reduce the amount of processing time needed. On top of that, we can easily extract other useful information from the collective knowledge such as the name of the player that scored the goal. This was previously a difficult task in video content analysis because of the lack of visually distinctive characteristics. Note that we did not only annotate the video content with objective concept information, but we also enriched the video content with subjective user information.

5. CONCLUSIONS AND FUTURE WORK

The lack of generalizability and the limited amount of detectable concepts are currently the main drawbacks of algorithms for automatic video annotation. In this Ph.D. proposal, we proposed a multi-modal fusion approach to overcome these limitations, making use of freely available collective knowledge in social media. To demonstrate the usefulness of the proposed approach, we successfully extracted goals and elaborate descriptions from broadcast soccer video sequences with limited effort.

In future research, we will further enhance the proposed annotation framework. One of our goals is to be able to annotate a full evening of broadcast video, with the broadcast video originating from several channels and containing several types of content. To that end, we will have to generalize the current annotation framework used. Further, compared to other external data sources, the power of collective knowledge lies in its user-generated nature. As such, in future research, we will design methods that are able to leverage collective knowledge for both the purpose of annotating content with a rich concept vocabulary and the purpose of enriching content with opinions and emotions.

6. ACKNOWLEDGMENTS

The research activities described in this paper were funded by Ghent University, iMinds, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

7. REFERENCES

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems*, 2010.

[2] J. Hannon, K. McCarthy, J. Lynch, and B. Smyth. Personalized and automatic social summarization of events in video. In *Proc. of the 16th international conference on Intelligent user interfaces*, 2011.

[3] J. Lanagan and A. F. Smeaton. Using twitter to detect and tag important events in live sports. *Artificial Intelligence*, 2011.

[4] V. Robu, H. Halpin, and H. Shepherd. Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Trans. Web*, 2009.

[5] D. Shamma, L. Kennedy, and E. Churchill. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events? *Horizon, In CSCW 2010*, 2010.

[6] X. Shi, Z. Yang, M. Toyoda, and M. Kitsuregawa. Harnessing the wisdom of crowds: video event detection based on synchronous comments. In *Proc. of the 20th international conference companion on World wide web*, 2011.

[7] G. van Oorschot, M. van Erp, and C. Dijkshoorn. Automatic extraction of soccer game events from twitter. In *Proc. of the Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*, 2012.

[8] L. Xie, L. Kennedy, S. fu Chang, A. Divakaran, H. Sun, and C. yung Lin. Layered dynamic mixture model for pattern discovery in asynchronous multi-modal streams. In *International Conference on Acoustic, Speech and Signal Processing*, 2005.

[9] C. Xu, Y.-F. Zhang, G. Zhu, Y. Rui, H. Lu, and Q. Huang. Using webcast text for semantic event detection in broadcast sports video. *Multimedia, IEEE Transactions on*, 2008.

[10] J. Yang and A. G. Hauptmann. (un)reliability of video concept detection. In *Proc. of the int. conference on Content-based image and video retrieval*, 2008.