Summarization and Presentation of Real-Life Events Using Community-Contributed Content

Manfred del Fabro and Laszlo Böszörmenyi

Institute of Information Technology, Klagenfurt University, Austria
{manfred,laszlo}@itec.aau.at

Abstract. We present an algorithm for the summarization of social events with community-contributed content from Flickr and YouTube. A clustering algorithm groups content related to the searched event. Date information, GPS coordinates, user ratings and visual features are used to select relevant photos and videos. The composed event summaries are presented with our video browser.

1 Introduction

Twenty years ago people were informed about a big social event, such as a royal wedding, essentially through a few, authorized, professional camera teams and journalists of printed press. Nowadays, a vast amount of additional photos, videos and corresponding metadata are uploaded to social platforms, such as Flickr and YouTube. If we use these social platforms to get an overview of a specific social event, we receive a - usually extremely long - list of photos or videos. However, long lists are not suited to get a good overview. A compact presentation of a predefined length, which gives us a summary of the event would be desirable. Such a summary should consist of content of good technical quality, high diversity and high coverage [3].

We present an algorithm that summarizes real-life events based on community-contributed content. In a summary photos and videos can be mixed up. The aim is to provide a rich view of the original event that consists of the different views of different people that witnessed the event.

2 Summarization and Presentation of Events

Our algorithm has six input parameters: (1) Search terms that are passed directly to Flickr and YouTube as text queries. (2) The number of streams to be shown in parallel. (3) The maximum duration of the event summary. (4) The name of the location of the event. (5,6) The dates of the lower and the upper bound of the timespan when the content must have been produced.

A summary may consist of more than a single sequence of photos and videos. While videos have a natural length we define a default duration – currently 7 seconds – for still images in the summary. In a single sequence an image needs

a rather short time, but as soon as more than one sequence is shown in parallel the viewers need more time to look at all photos shown in parallel.

The flow chart in Figure 1 illustrates the single steps of our event summarization algorithm. On Flickr we search for photos that were taken within the specified timespan. The YouTube API, unfortunately, does not allow to state a capturing or uploading date for the query. Therefore, we perform a post-processing step where we eliminate those videos that do not fit into the given timespan. The performance penalty for this is still acceptable, as we use only metadata for the post-processing.

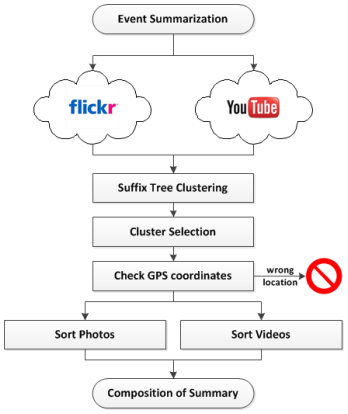

Fig. 1. Flow chart of algorithm

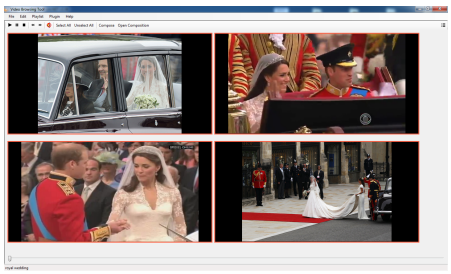

Fig. 2. Screenshot of event summary

We use a text suffix tree clustering algorithm [4] to group the content based on the textural descriptions. This algorithm has already successfully been applied to web document clustering. It is fast, separates relevant from irrelevant content and high quality clusters are produced even if only text snippets are available, which is indeed the case for the metadata of multimedia content. A dominant phrase is generated for each cluster. For the content selection we choose the largest cluster of which the dominant phrase includes the search terms of the query.

The textual descriptions of photos and videos are often misleading. We try to eliminate content produced in a wrong location by investigating the GPS coordinates of the content. The location indicated in textual form is translated into GPS coordinates using the Google Geocoding API[1] and matched against all photos and videos that have associated GPS data.

The selection of photos is based on the number of how frequently a photo has been viewed on Flickr. The selection of videos is based on the user ratings (up to 5 stars), the number of views and the number of "likes" a video has on YouTube. In general, the overall duration when photos are shown is approximately equal to the duration of the videos in the summary. This ratio is automatically adapted if the number of either the photos or the videos is too low. If the cluster selected

[1] Google Geocoding API: `http://code.google.com/apis/maps/documentation/geocoding/`

for the summary contains no videos or if the length of all videos exceeds the maximum duration, no videos are included.

To avoid redundancy we extract the Color and Edge Directivity Descriptor (CEDD) [1] from each candidate image to compare it with all images, which are already in the summary. If the distance to a photo in the summary is too low the candidate image will not be added. For videos the textual descriptions are sufficient to detect redundancy. Finally, when all photos and videos are selected we sort the whole content based on the timestamps.

Our own video browser [2] is used for the presentation of event summaries. The screenshot in Figure 2 shows a summary consisting of four parallel sequences. As parallel audio playback is not desirable, the audio stream of one of the videos is selected (either per default or by pointing at a sub-window).

To demonstrate our results we composed summaries of four well-known social events, which took place in the last three years: (1) the inauguration of Barack Obama, (2) the Royal Wedding of William and Kate, (3) the FIFA World Cup Final 2010 and (4) the UEFA Champions League Final 2011. Screen captures of the four composed event summaries are available online[2].

3 Conclusion and Future Work

In this paper we presented an algorithm for the summarization of real-life events based on community-contributed multimedia content. We used photos from Flickr and videos from YouTube to compose four summaries of events that attracted the attention of a lot of people.

The major future challenge is the temporal alignment of the content. The timestamps from the camera metadata are not sufficient. In our future work we are going to incorporate additional sources of information, like textual descriptions of the events, for the selection and the temporal alignment of content.

References

1. S. Chatzichristofis and Y. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Computer Vision Systems*, volume 5008 of *LNCS*, pages 312–322. Springer Berlin/Heidelberg, 2008.
2. M. del Fabro, K. Schoeffmann, and L. Böszörmenyi. Instant video browsing: A tool for fast non-sequential hierarchical video browsing. In *HCI in Work and Learning, Life and Leisure*, volume 6389 of *LNCS*, pages 443–446. Springer Berlin/Heidelberg, 2010.
3. P. Sinha, S. Mehrotra, and R. Jain. Summarization of personal photologs using multidimensional content and context. In *Proc. of the 1st ACM International Conference on Multimedia Retrieval*, pages 4:1–4:8, New York, NY, USA, 2011. ACM.
4. O. Zamir and O. Etzioni. Web document clustering: a feasibility demonstration. In *Proc. of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 46–54, New York, NY, USA, 1998. ACM.

[2] Demo videos: `http://soma.lakeside-labs.com/?page_id=279`