Summarization of Real-Life Events Based on Community-Contributed Content

Manfred del Fabro, Anita Sobe, Laszlo Böszörmenyi
Institute of Information Technology (ITEC)
AAU Klagenfurt University
Klagenfurt, Austria
{*manfred,anita,laszlo*}*@itec.aau.at*

***Abstract*—In this paper, we investigate whether community-contributed multimedia content can be used to make video summaries of social events. We implemented an event summarization algorithm that uses photos from Flickr and videos from YouTube to compose summaries of well-known society events, which took place in the last three years. The comparison with a manually obtained ground truth shows a good coverage of the most important situations of these events. We do not claim to produce the best summaries possible, which may be compared to the work of a human director, but we analyze what can be achieved with community-contributed content by now.**

***Keywords*-video summarization. event summarization. social media. real-life events. video retrieval. image retrieval. multimedia entertainment.**

I. Introduction

Twenty years ago, people were informed about a social event, such as a royal wedding, through a few, authorized, professional camera teams and journalists of printed press. Nowadays, a vast amount of additional photos, videos and some text, the latter mainly in form of metadata of the images, are uploaded to social platforms, such as Flickr and YouTube.

If we query these platforms to get informed about a certain social event, like the royal wedding of William and Kate in April 2011, we get a – usually extremely long – list of photos or videos. Even though the list is sorted corresponding to relevance, this is not a proper answer for such a question. We rather preferred to get a compact presentation of a predefined length, which gives us a summary, composed from the views of many people that have witnessed the event. This is not necessarily the best view, but the view that can be created based on the information people provided when uploading the content. Nevertheless, this view is usually very rich and contains a lot of interesting, even surprising elements. Of course, it may also contain garbage and even malicious content, but this is out of scope of this paper.

For this study, we only used social events related to entertainment. However, our approach is also applicable to other events, such as a traffic jam on a highway [23], seen by a number of drivers on the road, or a certain medical event, identified by a group of medical doctors in an arthroscopic surgery video [13].

This paper is organized as follows. Section II describes the related work, in Section III our event summarization algorithm is described in detail, in Section IV the experimental results are presented, and Section V concludes this paper and gives an outlook for future work.

II. Related Work

The summarization of multimedia content is the target of many research projects. Most of them focus on video abstraction and video summarization. Two extensive reviews of key-frame extraction and video summarization approaches are given in [15][22]. The presented algorithms summarize single videos with selected still images or with a short summary video. In our approach, we generate summaries that consist of content that comes from multiple sources.

One approach, which uses multiple videos as input for a summarization algorithm, is introduced in [10]. Videos of a whole basketball season in the USA and the corresponding metadata are used to create summary videos under different aspects, like summaries of the whole championship, of only one team or even of a single player. The authors only consider a single database for the content selection. Furthermore, professionally produced content from TV stations is used. No community-contributed annotations and ratings are available. In contrast, the system presented in this paper takes advantage of all context information that is provided by the community.

Not only the summarization of videos has been extensively studied, but the summarization of image collections has also been a target of research activities [19]. This work defines three aspects of an effective summary and formalizes models to optimize them: (1) quality of the content, (2) diversity of the content and (3) coverage of the whole collection. These aspects are also important for our summaries.

During the last few years, more and more research activities were focusing on real-life events in the context of multimedia data. In [26], a common event model for multimedia applications is proposed. Eight basic aspects are defined to describe an event, but also the relationship of an event to other events.

An event-based clustering algorithm is proposed in [14]. A layered clustering algorithm produces different clusters of videos, where each cluster represents one event.

 ISBN: 978-1-61208-195-3

In [12], information from online event directories is used to get metadata about an event, like the title or the geo information. With the help of this additional information the authors try to gather as many photos and videos as possible from Flickr and YouTube.

A visual-based method for retrieving events in photo collections of community-contributed contents is introduced in [21]. Based on a query image, an image collection is searched for similar photo records that may be of the same event.

In [25], an automatic remixing approach for community-contributed content from music concerts is presented. Users can record and upload videos during live events. Afterwards, the shared content is synchronized based on the creation timestamps and a master audio track is extracted from the single audio tracks of the synchronized videos. In the end, video remixes of a concert are automatically created based on automatically detected regions of interest.

The organization of tagged photo collections based on landmark and event detection is presented in [16]. Photos are arranged on their spatial closeness and their relatedness to events.

In [9], a joint content-event model is proposed, which allows an event-based indexing of videos instead of a concept-based one. A content model that describes videos in terms of scenes and shots is linked to an event model that defines different events and how they may be related to each other.

All these event-related approaches identify events in multimedia content or they index or cluster the content according to events. However, in this paper, we investigate aspects how community-contributed content is suited for the generation of visual summaries of social events, which to the best of our knowledge has not been done before.

A work that is not event-centric but that shows the power of utilizing community-contributed content is presented in [18]. Images of online photo collections are used to generate 3D views of famous places in the world where a lot of photos are taken. The introduced application allows an exploration of places based on the content of people that have really been there.

III. Event Summarization

A summary of a social event should consider the three aspects, how to build a summary [19]: (1) quality, (2) diversity and (3) coverage. (1) Photos and videos of poor visual quality should be not included into the summary. (2) Similar photos or videos should not be included more than once. (3) The resulting summary should cover the event as good as possible showing as many situations that occurred as possible.

As the quality aspect has been intensively studied, we concentrate in this paper on the two other aspects. During the generation of the summaries we focus on the maximum diversity of the content. Our summarization algorithm may not produce the best summary possible, but it creates a representation that emerges from most relevant and most popular contents related to a certain event of social media sharing platforms. In our evaluation we then investigate the coverage of that emerging view.

A. Summarization Algorithm

A summary is built according to search terms, specified by the user, such as: *Royal wedding of William and Kate*. First we cluster the content, based on the available textual descriptions. After that we filter wrongly located content based on GPS information. At last, we create a summary, from the remaining content. A flow chart, which illustrates these steps, is shown in Figure 1.

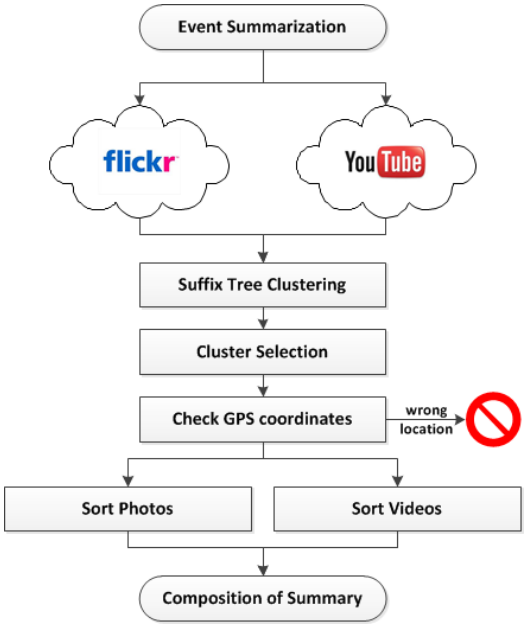

Figure 1. Flow chart of algorithm

The composition of an event summary is influenced by six parameters: (1) search terms to describe the event with keywords, (2) number of photos or videos to be shown in parallel, (3) maximum duration of the summary in seconds, (4) location, (5) start of the timespan the content must have been produced, and (6) end of the timespan the content must have been produced.

The search terms are passed to Flickr and YouTube as text queries. The more comprehensive the query, the better focused the retrieved content will be. Therefore, different queries lead to different summaries. The results of both platforms are sorted by relevance. We rely on the relevance calculations of both platforms and do not perform our own ones. This is the default sorting mode of both platforms. If people use the web interfaces of Flickr or YouTube, they also get the results sorted by relevance. A summary may

 ISBN: 978-1-61208-195-3

consist of more than a single sequence of photos and videos. Figure 2 shows a screenshot of an event summary, which consists of four parallel streams.

We only query for content that has been produced within the indicated timespan. We are well aware that the timestamps of the photos and videos may be wrong or even missing. We are going to pay attention to this fact in our evaluation. The queries do not return the photos and videos themselves, but only their metadata. For runtime reasons, we decided to limit the amount of Flickr results to 5000 per query. The YouTube API limits the amount of results to 1000 per query. The amount of photos and videos we consider for the summary generation is still much larger than a user would manually examine when clicking through Flickr and YouTube results. Therefore, we are of the opinion that this limitation is reasonable.

Figure 2. Screenshot of event summary

B. Clustering

In the next step, we cluster the photos and videos based on their textural descriptions. For that purpose, we use the text suffix tree clustering algorithm introduced in [27]. It has already successfully been applied to web document clustering and shows some interesting properties that can be exploited for our task.

The suffix tree clustering algorithm separates relevant from non-relevant content, even if only text snippets are available. Furthermore, the authors of the algorithm showed that it works fast. Therefore, it is well suited for multimedia content, as for photos and videos typically only short descriptions are available. For each photo and video retrieved from Flickr or YouTube, we extract title, description and tags. This information is the input for the clustering algorithm. At the end, we receive several clusters consisting of photos and videos. For each cluster, a summary in form of a *dominant phrase* is provided by the clustering algorithm. For the content selection, we choose the largest cluster of which the dominant phrase includes the search terms of our query.

C. Content selection and composition of event summaries

Photos and videos often have misleading descriptions regarding their location. We try to overcome this problem by investigating the GPS coordinates of the content. The location indicated in textual form is translated in GPS coordinates using the Google Geocoding API [2]. Using the retrieved GPS coordinates and the level of detail (country, region, city or street) we are able to eliminate content that has been produced in a wrong place. If ambiguities are possible (e.g., Paris, France and Paris, Texas), the location must be specified precisely. Otherwise wrong content may be included in the summary.

The selection of photos for the summary 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. For each event summary, we select content in such a way that the amount of time when photos are shown and the amount of time when videos are played are approximately equal. While videos have a natural length we define a default duration of 7 seconds for still images in the summary. In a single sequence this may be too long to show a single image, but as soon as more than one sequence is shown in parallel the viewers need more time to look at all photos. For example, for a video with a duration of 28 seconds we add four photos to the summary. This ratio is automatically adapted if the number of either the photos or the videos is too low. It may happen that no videos are included in a summary, because the selected cluster does not contain videos at all or the length of the contained videos exceeds the maximum duration of the summary.

One important aspect of the summarization of content is to avoid redundancy [22]. We rely on visual image features to identify redundant photos. Each image selected as a candidate for a summary is matched against all other photos that are already in the summary. If the visual similarity to a photo in the summary is too high, the candidate image will not be added. For the estimation of the visual similarity, we extract the Color and Edge Directivity Descriptor (CEDD) [7] from each photo. The CEDD can be extracted fast and it showed good results in an evaluation of different image features for video summarization [11].

Finally, when all photos and videos are selected we sort the whole content based on their creation timestamps. With this simple approach we want to investigate how good timestamps are suited to make a temporal alignment of the content.

D. Summary format and presentation

In the resulting event summary, the videos are played first and then slide shows of the photos are shown. We think that the viewers get a good impression and an overview by watching the videos first, while photos are better suited to cover certain aspects in detail that the videos may miss.

 ISBN: 978-1-61208-195-3

Table I
DETAILS ABOUT COMMUNITY-CONTRIBUTED DATA RELATED TO CERTAIN SOCIAL EVENTS

Search terms	inauguration obama	royal wedding	fifa world cup final 2010	champions league final 2011
Flickr results	59643	47372	2535	1529
YouTube results	15800	52500	547000	186000
Photos/Users selected	1062/182	1516/343	668/81	161/22
Videos/Users selected	1/1	211/211	114/90	83/72
Photos with GPS	333	437	333	42
Videos with GPS	0	7	7	0
Wrong location	81	211	160	1
Photos/Users in summary	168/51	73/28	81/17	83/14
Videos/Users in summary	0/0	5/5	4/4	5/5

We present the generated summaries in our own Video Browser [8], which is depicted in Figure 2. This video browser allows showing of several videos and photos in parallel. The audio playback is selected from one of the presented videos by default or by mouse-over on one of the videos.

To organize the temporal presentation the player interprets a formalism called Video Notation (ViNo) [20]. ViNo is a multipurpose multimedia language, which we use to define the presentations in a short and flexible way. Each event summary is a ViNo expression, which consists of a sequence of videos or photos shown in parallel. E.g., the presentation of four videos as shown in Figure 2 can be expressed in ViNo as $[u1||u2]|[u3||u4]$, where we assume that u*i* is the identifier of a video and $||$ means parallel presentation. Each line is grouped by squared brackets.

IV. EVALUATION

We chose four well-known social events that took place in the last three years for the evaluation: (1) the inauguration of Barack Obama [5], (2) the Royal Wedding of William and Kate [6], (3) the FIFA World Cup Final 2010 [3] and (4) the UEFA Champions League Final 2011 [4]. All four events took place on one single day, were attended by several thousands of people and attracted the attention of millions of people around the world.

The same algorithm was used for all four summaries. We did not tune it according to the events. All event summaries in our evaluation consist of 4 parallel streams and have a maximum duration of 5 minutes. The timespan we used for our queries starts with the day the event took place and ends one month after that. Other investigations showed that even a time interval of 7 days is sufficient [12]. Screen captures of the four composed event summaries are available online [1].

Table I lists the *Search terms* that were used as input for the summary generation and gives details about the retrieved content. We tried to use as few search terms as possible to describe the events, because people also tend to use only a few terms when searching for multimedia content online [24].

The same query, which is used for the summary generation, has also been used to query the Flickr (*Flickr results*) and the YouTube (*YouTube results*) website to get a first impression of the available content. For the first two queries much more photos can be retrieved from Flickr than for the two soccer matches. The reason for that is that more specific text queries were used for the two soccer matches consisting of 4 and 5 terms, compared with only 2 terms for the first two queries. The more specific a query is the less results are returned from Flickr. Interestingly, for the two soccer matches a huge amount of videos is available. A closer examination shows that people played these matches also on their gaming consoles and published videos of that computer games online.

The event summary algorithm originally included the 5000 most relevant Flickr and the 1000 most relevant YouTube results. Finally, even a smaller subset – as produced by the clustering – is used for the content selection. The rows *Photos/Users selected* and *Videos/Users selected* list how many photos and videos were included in the final cluster for the summary generation and how many distinct users uploaded these contents. It can be seen that several photos are selected from each included Flickr uploader, while in most cases the included YouTube videos have different users.

In the created summaries 3 to 6 photos of a single uploader (*Photos/Users in summary*) are included. Each video in these summaries (*Videos/Users in summary*) has a single uploader. The summary of the inauguration of Barack Obama only consists of photos. The cluster selected for the

 ISBN: 978-1-61208-195-3

summary only contains one video of his oath, but its length exceeds the maximum duration of the summary. In general, these summaries include content from a variety of users, thus these summaries are really conveying a broad view of people that witnessed the selected events.

The retrieved data shows that the available GPS data provide only a strongly limited support to estimate the location where the content was produced. For only 25 – 50 % of the selected photos (*Photos with GPS*) are the GPS data available and videos (*Videos with GPS*) hardly having this data associated at all. Nevertheless, many photos could be filtered that were taken in a wrong location. The relatively high amount of photos excluded due to wrong semantic location (*Wrong location*) can be easily explained. The events chosen for the summaries were broadcasted all over the world. The excluded content was produced by people somewhere else on the world. In most cases people celebrated parties to follow the original event in a group on TV. The content produced at those parties was annotated with textual descriptions related to the original event. Therefore, it was initially included in the results sets retrieved by Flickr and YouTube.

The coverage of the created summaries is compared against a manually obtained ground truth. The most important *situations* of the chosen events were figured out with the help of Wikipedia articles [3][4][5][6]. For each event a corresponding set of situations was identified. A situation may be a temporal happening, such as *exchange of the rings*, a location, such as the Westminster Abbey or even persons, such as *Prince Harry*. Table II lists the identified situations for all four events. Further information about these situations can be obtained from the Wikipedia articles. Later in this paper, we refer to these four lists of situations when the evaluation of the coverage of the generated summaries is presented.

We decided to rely on Wikipedia, because it is difficult to find an objective evaluation metric for the quality of summaries. Summaries are always somehow based on subjective opinions as [17] showed. Wikipedia articles usually have several authors, who perform discussions and have to agree on the text of the article. Therefore, Wikipedia articles convey the common opinion of a crowd of people. We take advantage of that common opinion to get a more objective ground truth for the evaluation of the coverage of the generated event summaries.

We compared our event summaries with a standard web search on Flickr and YouTube. As the evaluated summaries have a duration of 5 minutes, we limited the number of Flickr and YouTube results to amounts that could approximately be browsed in that time span. The first 120 photos from Flickr and the first 20 videos from YouTube are investigated for each query. If we compare the coverage of the generated summaries with the Flickr and YouTube results in the following parts of this evaluation, we always refer to result sets of that size (indicated by *Flickr* resp. *YouTube* in the following diagrams).

Table II
INTERESTING SITUATIONS OF THE FOUR SOCIAL EVENTS

Inauguration Obama	**Royal Wedding**
1. United States Capitol	1. Westminster Abbey
2. Music live performances	2. Bride (Kate)
3. Invocation by pastor	3. Groom (William)
4. Aretha Franklin singing	4. Pippa Middleton
5. Oath of Vice President	5. Prince Harry
6. Oath of Barack Obama	6. Queen Elisabeth II.
7. Inaugural address	7. Young bridesmaids
8. Prayers	8. Pageboys
9. Departure of former president	9. Arrival of Kate
10. Signing of first orders	10. Exchange of rings
11. Luncheon	11. Lesson
12. Parade	12. Sermon
13. Inauguration balls	13. Leaving Westminster Abbey
14. National prayer service	14. Return to palace in coach
15. Oath of office	15. Lunchtime reception
	16. Appearing on balcony
	17. Harpist performance
	18. William & Kate leaving with car
	19. Private dinner
	20. Wedding cake
	21. Merchandise
	22. Broadcasting
World Cup	**Champions League**
1. Soccer City Stadium	1. Wembley Stadium
2. de Jong's kick against Alonso	2. Chance Hernandez (ManU)
3. Chance Robben (NED)	3. Chance Villa (Barca)
4. Chance Sneijder (NED)	4. Chance Villa (Barca)
5. Chance Ramos (ESP)	5. Goal Pedro (Barca)
6. Red card Heitinga (NED)	6. Goal Rooney (ManU)
7. Goal Iniesta (ESP)	7. Chance Messi (Barca)
8. Award ceremony	8. Chance Messi (Barca)
	9. Goal Messi (Barca)
	10. Chance Messi (Barca)
	11. Chance Xavi (Barca)
	12. Goal Villa (Barca)
	13. Chance Rooney (ManU)
	14. Chance Nani (ManU)
	15. Award ceremony

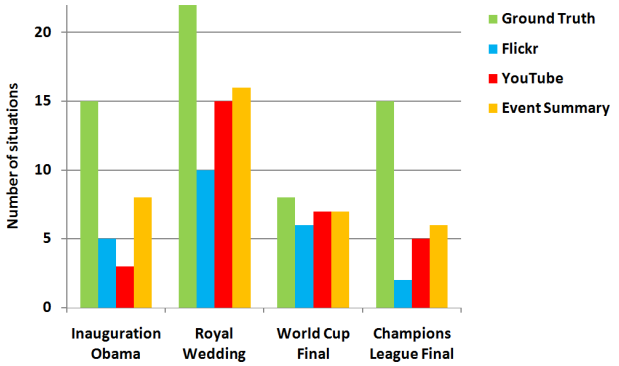

Figure 3. Comparison of situations found

The results are shown in Figure 3. In all cases, the first Flickr results only include few situations of interest. The reason for that is that people tend to photograph themselves

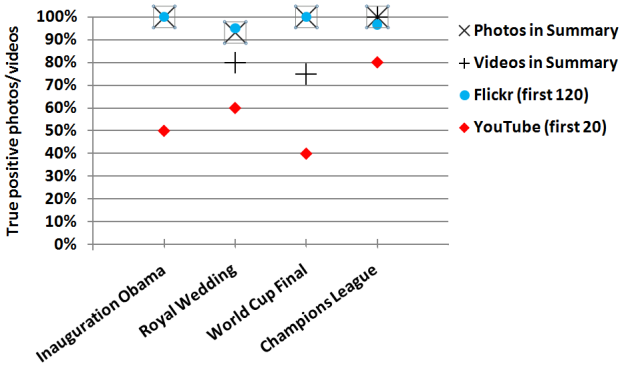

Figure 4. Amount of true positive photos or videos

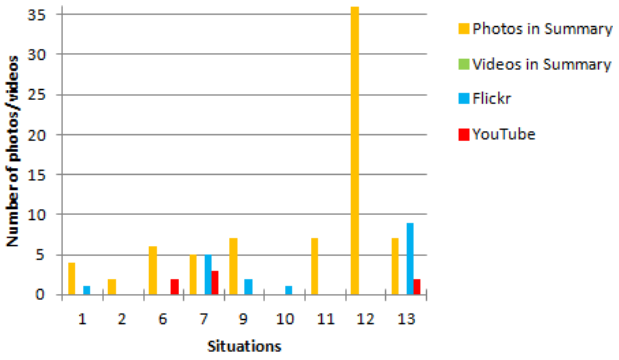

Figure 5. Inauguration Obama

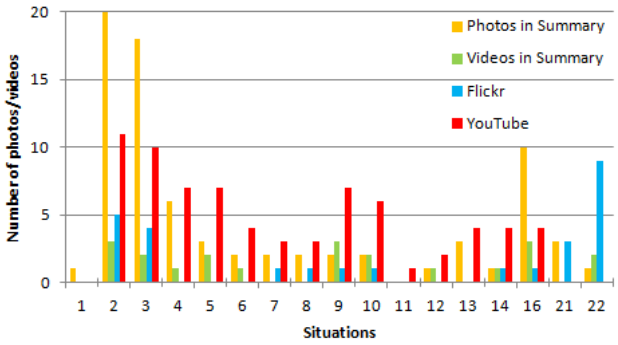

Figure 6. Royal Wedding

when visiting an event. Therefore, a lot of images show visitors of the event and only few photos show situations as they were identified based on the Wikipedia entries. Except for the inauguration of Obama the YouTube results show more interesting situations than the Flickr results. The event summarization algorithm shows in all cases the best performance. It includes as much situations as Flickr or YouTube or even more.

If content is examined regardless of the searched situations, it can be recognized that the precision of the Flickr results is high. They include a high amount of content that is related to the searched events. Figure 4 shows the percentage of true positive photos and videos in the Flickr and YouTube results as well as in the event summaries. For the latter, we distinguish between photos and videos. A photo or video is regarded to be a true positive if it is somehow related to the event. The Flickr results contain a lot of true positives, which also has a positive effect on the photos in the summaries. Except for the Champions League final the YouTube results have a lot of false positives, although only the 20 most relevant results returned by YouTube are considered. The event summarization algorithm also includes false positives in the summaries, but the ratio of true positives is much better than the one of the YouTube results. This is an effect of the suffix tree clustering of the content. As the biggest cluster is chosen, which is related to the query, it is more likely that this cluster includes relevant content. Note that false positives include photos and videos, which are not wrong, rather strange. For example, if some people record the movements of the police at the royal wedding (as they did indeed), this is topic for a non-technical discussion, whether or not these images are misplaced.

The comparison of the coverage shows that quite a lot of the defined situations of interest are not included in the summaries as well as in the Flickr or YouTube results. Therefore, we want to take a closer look at the situations found. Figure 5 shows the situations detected for the inauguration of Barack Obama. It can be noticed that a lot of photos are showing the parade (situation no. 12) after the inauguration. That was somehow expected, because the parade was watched by a lot of people along the track and thus a lot of photos have been made. For the other situations it can be stated that people especially took photos of the highlights, like the oath of Obama (6), his inaugural address (7) or the departure of the former president Bush (9). Also the society events like the luncheon (11) and the balls (13) seem to attract people. The oath of the Vice-President (5), prayers (8) or events that took place in the office of Obama, like the signing of the first orders (10) or his second oath (15) are not covered by the content we received from Flickr and YouTube.

Figure 6 shows the identified situations of the royal wedding in detail. As it can be seen the involved people like Kate (2), William (3), Pippa (4), Prince Harry (5) or the Queen (6) get a lot of attention. Also the appearing on the balcony (16) or situations that took place in the streets or in front of the church (9, 13 and 14) are included often. The reason is again that for public situations a lot of content is produced, while for private ones like the family celebrations (15) or the private dinner (19) in the evening nothing can be found.

We also wanted to investigate events where the interesting situations may be clearer. Therefore, we decided to investigate event summaries of two soccer games that attracted

 ISBN: 978-1-61208-195-3

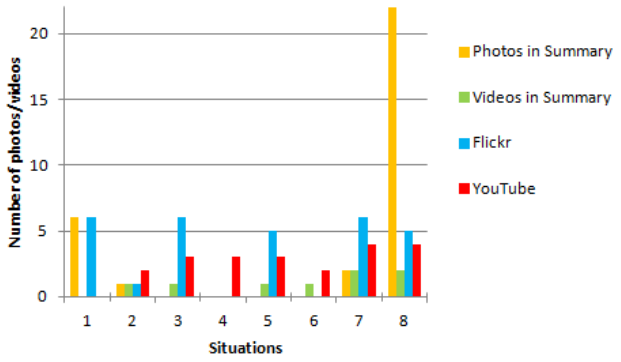

Figure 7. FIFA World Cup Final 2010

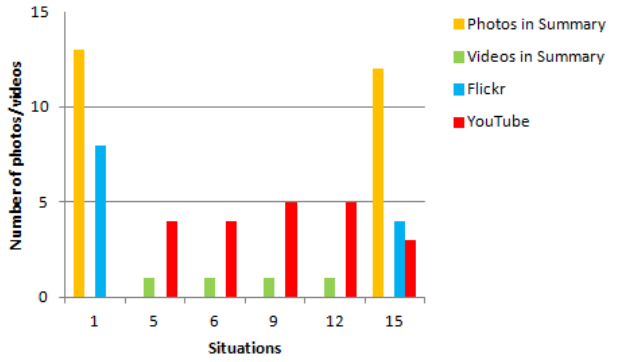

Figure 8. UEFA Champions League Final 2011

the attention of millions of people around the world. The identified situations for the two games are shown in Figure 7 and Figure 8. For both games it can be stated that all goals are identified by our summaries, but nearly all chances that did not result in goals are missed. In addition to the goals both summaries also include a lot of situations showing the venue and the award ceremonies of the winning teams.

Regarding the temporal alignment of the content we must state that the timestamps of the content are not sufficient for good ordering of the content. By simply watching the generated summaries it can be seen that the content is mixed up temporarily in all summaries. It seems that people do not care about their cameras having correct date and time settings. Nevertheless, this could change, if people notice in the future that innovative tools can make good use of this information.

V. Conclusion

In this paper, we presented an algorithm for the summarization of real-life events based on community-contributed multimedia content. We composed four summaries of events that attracted a lot of people during the last three years using photos from Flickr and videos from YouTube. We evaluated the coverage of our summaries by comparing them with Wikipedia articles that report about the corresponding events. This innovative evaluation technique allows us to identify the important happenings of social events without doing manual observations of these events, but by relying on the common opinion of a group of people that created and edited the corresponding articles. Furthermore, we investigated some characteristics of community-contributed content with respect to event summarization. The composed summaries show a good coverage of interesting situations that happened during the selected events. Next, we plan to perform user studies to investigate how the quality of our summaries is perceived by people that watch them.

There are still several open questions that remain in this rather new topic, like the correct temporal alignment of content or the identification of malicious content. In our future work we are also going to incorporate additional sources of information, like textual descriptions of the events, for the temporal alignment as well as for the selection of content. Furthermore, we are going to make investigations regarding the sensitivity of our algorithm to be able to state which results can be achieved under which circumstances.

Further investigations have to be done on events that last longer than one day (e.g., the whole FIFA World Championship), events that have many parallel sub-events (e.g., Olympic Games), and small events (which only attract the attention of a small audience). But not only events that are related to entertainment are of interest. The presented approach can also be applied to spontaneous real-life events like a traffic jam on a motorway or a catastrophe scenario like a heavy earthquake. If summaries of such events are constructed from the content that involved people or witnesses have captured, emergency response teams may profit from that information and may be steered and coordinated in a better way.

Acknowledgment

This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria and funding from the European Regional Development Fund and the Carinthian Economic Promotion Fund (KWF) under grant KWF-20214 17097 24774 and grant KWF-20214 22573 33955.

References

[1] Demo videos of event summaries. http://soma.lakeside-labs.com/?page_id=279. [2012-02-09].

[2] Google geocoding api. http://code.google.com/apis/maps/documentation/geocoding/. [2012-02-09].

[3] Wikipedia: 2010 FIFA world cup final. http://en.wikipedia.org/w/index.php?title=2010_FIFA_World_Cup_Final&oldid=439386816. [2012-02-09 (Permalink)].

[4] Wikipedia: 2011 UEFA champions league final. http://en.wikipedia.org/w/index.php?title=2011_UEFA_Champions_League_Final&oldid=440623020. [2012-02-09 (Permalink)].

 ISBN: 978-1-61208-195-3

[5] Wikipedia: Inauguration of barack obama. http://en.wikipedia.org/w/index.php?title=Inauguration_of_Barack_Obama&oldid=439374433. [2012-02-09 (Permalink)].

[6] Wikipedia: Wedding of prince william and catherine middleton. http://en.wikipedia.org/w/index.php?title=Wedding_of_Prince_William_and_Catherine_Middleton&oldid=440475841. [2012-02-09 (Permalink)].

[7] S. Chatzichristofis and Y. Boutalis. CEDD: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In A. Gasteratos, M. Vincze, and J. Tsotsos, editors, *Computer Vision Systems*, volume 5008 of *Lecture Notes in Computer Science*, pages 312–322. Springer Berlin / Heidelberg, 2008.

[8] M. del Fabro, K. Schoeffmann, and L. Böszörmenyi. Instant video browsing: A tool for fast non-sequential hierarchical video browsing. In G. Leitner, M. Hitz, and A. Holzinger, editors, *HCI in Work and Learning, Life and Leisure*, volume 6389 of *Lecture Notes in Computer Science*, pages 443–446. Springer Berlin / Heidelberg, 2010.

[9] N. Gkalelis, V. Mezaris, and I. Kompatsiaris. Automatic event-based indexing of multimedia content using a joint content-event model. In *Proceedings of the 2nd ACM international workshop on Events in multimedia*, EiMM '10, pages 15–20, New York, NY, USA, 2010. ACM.

[10] R. Kaiser, M. Hausenblas, and M. Umgeher. Metadata-driven interactive web video assembly. *Multimedia Tools and Applications*, 41:437–467, 2009.

[11] M. Kogler, M. del Fabro, M. Lux, K. Schoeffmann, and L. Böszörmenyi. Global vs. local feature in video summarization: Experimental results. In *Proceedings of the 10th International Workshop of the Multimedia Metadata Community on Semantic Multimedia Database Technologies (SeMuDaTe09) in conjunction with the 4th International Conference on Semantic and Digital Media Technologies (SAMT 2009).*

[12] X. Liu, R. Troncy, and B. Huet. Finding media illustrating events. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 58:1–58:8, New York, NY, USA, 2011. ACM.

[13] M. Lux, O. Marques, K. Schöffmann, L. Böszörmenyi, and G. Lajtai. A novel tool for summarization of arthroscopic videos. *Multimedia Tools and Applications*, 46:521–544, 2010.

[14] J. Makkonen, R. Kerminen, I. D. D. Curcio, S. Mate, and A. Visa. Detecting events by clustering videos from large media databases. In *Proceedings of the 2nd ACM international workshop on Events in multimedia*, EiMM '10, pages 9–14, New York, NY, USA, 2010. ACM.

[15] A. G. Money and H. Agius. Video summarisation: A conceptual framework and survey of the state of the art. *J. Vis. Comun. Image Represent.*, 19(2):121–143, February 2008.

[16] S. Papadopoulos, C. Zigkolis, Y. Kompatsiaris, and A. Vakali. Cluster-based landmark and event detection for tagged photo collections. *Multimedia, IEEE*, 18(1):52 –63, jan. 2011.

[17] G. J. Rath, A. Resnick, and T. R. Savage. The formation of abstracts by the selection of sentences. part i. sentence selection by men and machines. *American Documentation*, 12(2):139–141, 1961.

[18] I. Simon, N. Snavely, and S. M. Seitz. Scene summarization for online image collections. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct. 2007.

[19] P. Sinha, S. Mehrotra, and R. Jain. Summarization of personal photologs using multidimensional content and context. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 4:1–4:8, New York, NY, USA, 2011. ACM.

[20] A. Sobe, L. Böszörmenyi, and M. Taschwer. Video Notation (ViNo): A Formalism for Describing and Evaluating Non-sequential Multimedia Access. *International Journal on Advances in Software*, 3(1 & 2):19–30, sep 2010.

[21] M. R. Trad, A. Joly, and N. Boujemaa. Large scale visual-based event matching. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ICMR '11, pages 53:1–53:7, New York, NY, USA, 2011. ACM.

[22] B. T. Truong and S. Venkatesh. Video abstraction: A systematic review and classification. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(1):3+, 2007.

[23] R. Tusch, A. Fuchs, H. Gutmann, M. Kogler, J. Köpke, L. Böszörmenyi, M. Harrer, and T. Mariacher. A multimedia-centric quality assurance system for traffic messages. In J. Düh, H. Hufnagl, E. Juritsch, R. Pfliegl, H.-K. Schimany, and H. Schönegger, editors, *Data and Mobility*, volume 81 of *Advances in Intelligent and Soft Computing*, pages 1–13. Springer Berlin / Heidelberg, 2010.

[24] R. van Zwol, B. Sigurbjornsson, R. Adapala, L. Garcia Pueyo, A. Katiyar, K. Kurapati, M. Muralidharan, S. Muthu, V. Murdock, P. Ng, A. Ramani, A. Sahai, S. T. Sathish, H. Vasudev, and U. Vuyyuru. Faceted exploration of image search results. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 961–970, New York, NY, USA, 2010. ACM.

[25] S. Vihavainen, S. Mate, L. Seppälä, F. Cricri, and I. D. Curcio. We want more: human-computer collaboration in mobile social video remixing of music concerts. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, pages 287–296, New York, NY, USA, 2011. ACM.

[26] U. Westermann and R. Jain. Toward a common event model for multimedia applications. *IEEE Multimedia*, 14(1):19–29, 2007.

[27] O. Zamir and O. Etzioni. Web document clustering: a feasibility demonstration. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '98, pages 46–54, New York, NY, USA, 1998. ACM.

 ISBN: 978-1-61208-195-3