Key Event Detection in Video using ASR and Visual Data

Niraj Shrestha **Aparna N. Venkitasubramanian** **Marie-Francine Moens**
KU Leuven, Belgium
{niraj.shrestha, Aparna.NuraniVenkitasubramanian,
Marie-Francine.Moens}@cs.kuleuven.be

Abstract

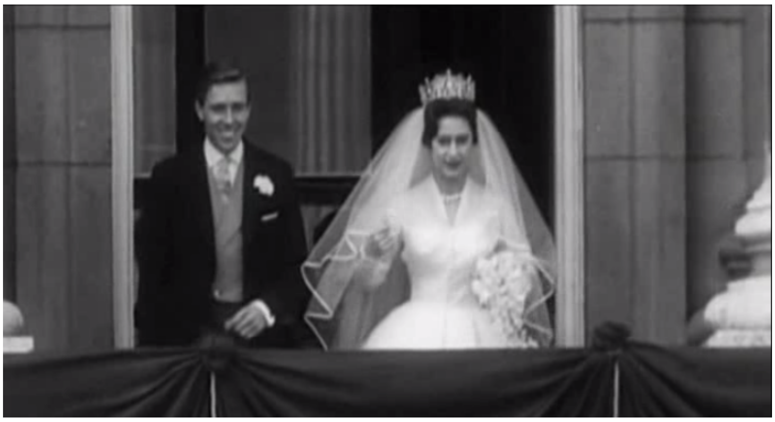
Multimedia data grow day by day which makes it necessary to index them automatically and efficiently for fast retrieval, and more precisely to automatically index them with key events. In this paper, we present preliminary work on key event detection in British royal wedding videos using automatic speech recognition (ASR) and visual data. The system first automatically acquires key events of royal weddings from an external corpus such as Wikipedia, and then identifies those events in the ASR data. The system also models name and face alignment to identify the persons involved in the wedding events. We compare the results obtained with the ASR output with results obtained with subtitles. The error is only slightly higher when using ASR output in the detection of key events and their participants in the wedding videos compared to the results obtained with subtitles.

1 Introduction

With the increase of multimedia data widely available on the Web and in social media, it becomes necessary to automatically index the multimedia resources with key events for information search and mining. For instance, it is not possible to manually index all the frames of a video. Automatically indexing multimedia data with key events makes the retrieval and mining effective and efficient.

Event detection is an important and current research problem in the field of multimedia information retrieval. Most of the event detection in video is done by analyzing the visual features using manually transcribed data. In this paper, we propose key event detection in British royal wedding videos using automatic speech recognition (ASR) data and where possible also to recognize the actors involved in the recognized events using visual and textual data. An event is something that happens at a certain moment in time and at a certain location possibly involving different actors. Events can be quite specific as in this case the key events are the typical events that make up a royal wedding scenario. For example, events like 'design of cake/dress/bouquet', 'couple heading to Buckingham palace', 'appearing on balcony' etc. are key events in British royal wedding video. Figure 1 shows an example of a frame containing an event with its actors, together with the associated subtitle and ASR output. While most works in this domain have focussed on clean textual content such as manual transcripts or subtitles, which are difficult to acquire, we use the output of an ASR system. While the event detection and name-face alignment problem by itself is already quite difficult, the nature of the ASR text adds an additional complexity. ASR data is noisy and inaccurate, it does not contain some parts of the actual spoken text, and does not contain sentence boundaries. Figure 2 illustrates this problem. For the key events, the system first acquires the necessary knowledge from external corpora - in our case Wikipedia articles associated with royal weddings. Then the system identifies the key events in the ASR data. The system also models name and face alignment to identify the persons involved in the wedding events. We perform named entity recognition in the text associated with a window of frames to first generate a noisy label for the faces occurring in the frames and this rough alignment is refined using an Expectation-Maximization (EM) algorithm. We compare the results obtained with the ASR output with results obtained with subtitles. The error is only slightly

Proceedings of the 25th International Conference on Computational Linguistics, pages 46–53,
Dublin, Ireland, August 23-29 2014.



Sub-title: ”Outside, fully 150,000 people with unbounded enthusiasm acclaimed Princess Margaret and her husband when they appeared on the balcony.. . . ”
ASR: ”outside only a hundred and 50 people on TV and using it as a . . . ”

Figure 1: An example of a frame containing an event with associated subtitle and ASR output

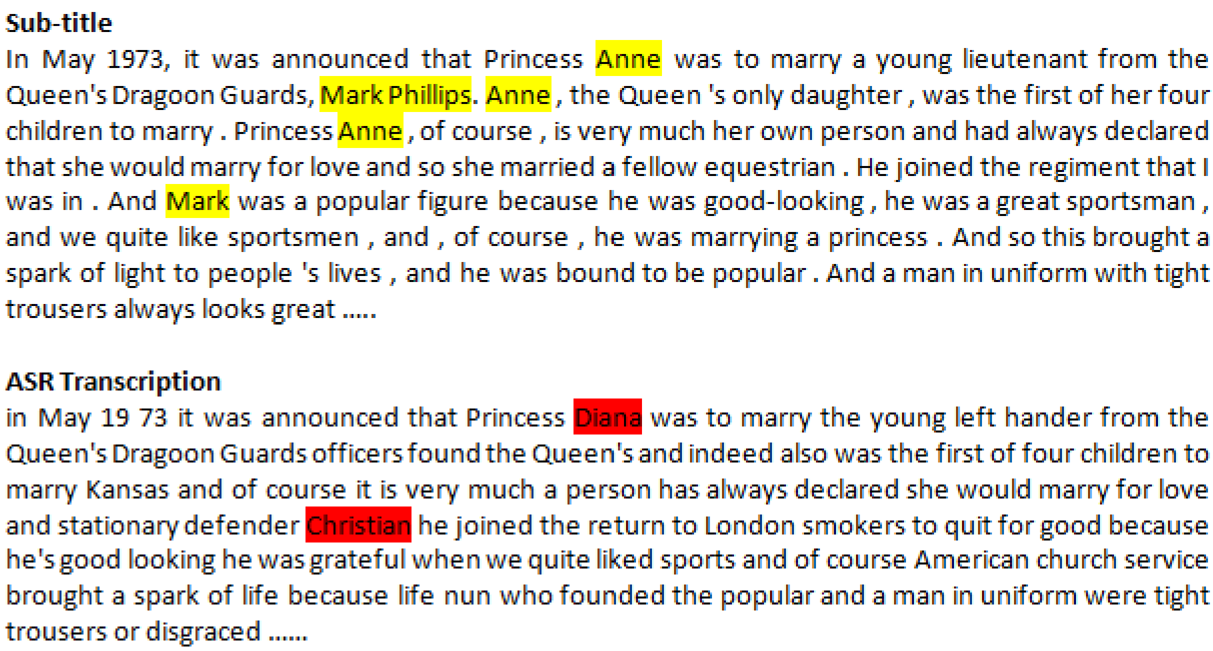
higher when using ASR output in the detection of key events and their participants in the wedding videos compared to the results obtained with subtitles. The methodology that we propose can be applied for the detection of many different types of video events.

2 Related work

Event detection has some relationship with Topic Detection and Tracking (TDT) and with concept detection. TDT regards the detection and tracking over time of the main event of a news story and is a challenging problem in the field of text and visual analysis. Although widely studied in text (Allan, 2002), (Allan et al., 2005), (Mei and Zhai, 2005), (Wang et al., 2007), (Zhao et al., 2007), topic detection in video is still not well studied. An event in this context is usually broader in scope than the events we want to recognize in wedding videos in this paper. In the multimedia research community, most of the works focus on concept detection like in (Liu et al., 2008), (Yang et al., 2007), (Snoek et al., 2006) rather than event detection. A concept detection task is different from event detection as a concept can be defined as any object or specific configuration of objects. Any frame then can be labelled with some concept descriptor (e.g., church, cake, etc.). While in an event, there is a start and end time in between which something happens, and in video, an event is represented by a sequence of frames.

Event detection is a challenging problem which is not well studied. Only few event detection systems that process video exist. They recognize events such as goal, run, tackle in a soccer game, or recognize specific actions in news video (e.g., meeting of two well-known people) or in a surveillance video (e.g., unusual event). Event detection in video is often related to sports like basketball (Saur et al., 1997), soccer (Yow et al., 1995) and baseball (Kawashima et al., 1998) (Rui et al., 2000). (Wang et al., 2008) developed a model based on a multi-resolution, multi-source and multi-modal bootstrapping framework that exploits knowledge of sub-domains for concept detection in news video. (Adam et al., 2008) developed an algorithm based on multiple local monitors which collect low-level statistics to detect certain types of unusual events in surveillance video.

Most of these works rely only on visual analysis (e.g., detection of certain motion patterns) to identify events in video and the event detection is performed with a supervised learning method, where a model is trained on manually annotated examples of known events. In this paper, we propose a novel idea in which the system learns events from an external corpus like Wikipedia and identifies those events in the ASR or subtitle data of the video. In addition, we identify the persons involved in an event based on the analysis of visual and textual data.



Sub-title

In May 1973, it was announced that Princess Anne was to marry a young lieutenant from the Queen's Dragoon Guards, Mark Phillips. Anne , the Queen 's only daughter , was the first of her four children to marry . Princess Anne , of course , is very much her own person and had always declared that she would marry for love and so she married a fellow equestrian . He joined the regiment that I was in . And Mark was a popular figure because he was good-looking , he was a great sportsman , and we quite like sportsmen , and , of course , he was marrying a princess . And so this brought a spark of light to people 's lives , and he was bound to be popular . And a man in uniform with tight trousers always looks great .....

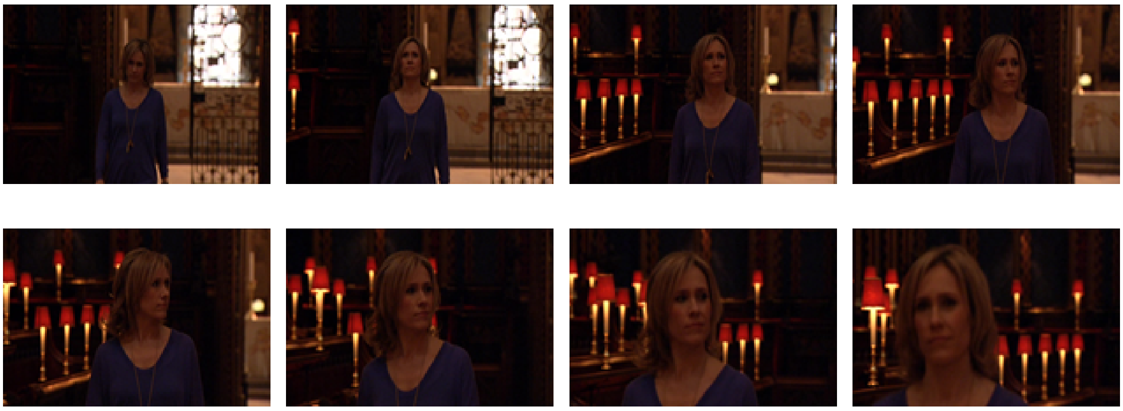
ASR Transcription

in May 19 73 it was announced that Princess Diana was to marry the young left hander from the Queen's Dragoon Guards officers found the Queen's and indeed also was the first of four children to marry Kansas and of course it is very much a person has always declared she would marry for love and stationary defender Christian he joined the return to London smokers to quit for good because he's good looking he was grateful when we quite liked sports and of course American church service brought a spark of life because life nun who founded the popular and a man in uniform were tight trousers or disgraced ......

Figure 2: An example showing sub-title vs. ASR data

3 Methodology

The main objective of this work is to identify and index key events in videos using ASR data along with key actors involved in the event. We start by identifying key events related to a certain domain, using external corpora. In addition, the proposed method involves pre-processing of the textual and visual data.



At 11.30, Elizabeth entered the abbey on her father's arm, but they did not head straight down the aisle as expected.

Figure 3: Sequence of frames showing the anchor talking about an event of the wedding, but there is no visual appearance of the event.

3.1 Acquiring background knowledge

Our approach for identifying key events in weddings exploits external text corpora. We use two corpora:

1. A genre-specific corpus: a set of pages specific to the topic, for example, from Wikipedia - to identify events associated with the topic.

2. A generic corpus, used to weigh the events identified in the genre-specific corpus.

The process is as follows. We first collect content from Wikipedia articles relevant for Britain's royal weddings[1] in the form of nine documents. These articles include both pages related to weddings, such as these of Diana and Charles, that were mentioned in our test videos as well as pages about other British royal weddings not shown in the videos, such as the wedding of Kate and William. This set of articles formed our wedding corpus for learning typical wedding events. The generic corpus is formed by all English Wikipedia articles.

From each document of this corpus we extract events together with their arguments (subject and object arguments) using a state-of-the-art event annotator[2]. This tool uses linguistic features such as the results of a dependency parse of a sentence, to detect the events and their arguments in a text. Next, we use a data mining technique to find frequent word patterns that signal the event and its arguments in the wedding articles, we keep each event that has sufficient support in the wedding articles and weigh it by a factor that is inversely proportional to the frequency of this event in the more general corpus. We keep the N highest weighted events from the obtained ranked list, where N is determined by whether we want to keep the most common wedding events or include also more rare events. The list obtained has items such as 'to announce engagement', 'to make dress', 'to make cake' etc, which are typical for weddings. We report here on preliminary work and acknowledge that the methodology can be largely refined.

3.2 Detecting person names

In royal wedding videos, there are many persons who appear in the video like anchor, interviewee, the persons married or to be married, the dress designer, the bouquet designer, the cake maker, the friends etc. As in this preliminary work we are only interested in the brides and bridegrooms (which are also the most important persons when indexing the video) we use a gazetteer with their names for recognizing the names in the texts.

3.3 Detecting the faces of persons

In the video key frames are extracted at the rate of 1 frame per second using (ffmpeg, 2012), which ensures that no faces appearing in the video are omitted. To detect the faces in the video, a face detector tool from (Bradski, 2000) is used. Next, we extract the features from the faces detected in the video. Although there are several dedicated facial feature extraction methods such as (Finkel et al., 2005),(Strehl and Ghosh, 2003), in this implementation, we use a simple bag-of-visual-words model (Csurka et al., 2004).

Once feature vectors are built, clustering of the bounding boxes of the detected faces is performed. Each object is, then, compared to the cluster centers obtained and is replaced with the closest center. The clustering is done using Elkan's k-means algorithm (Jain and Obermayer, 2010) which produces the same results as the regular k-means algorithm, but is computationally more efficient. This accelerated algorithm eliminates some distance calculations by applying the triangle inequality and by keeping track of lower and upper bounds for distances between points and centers. This algorithm, however, needs the number k of clusters present in the data. Since we are primarily interested in the brides and bridegrooms and since there are seven weddings shown in the video, we experiment with values of k equal to 7*2 = 14. Although this approach very likely introduces errors in the clustering as we do not know beforehand how many persons apart from the couple appear in the chosen key frames, it showed to be a better strategy than trying to align all persons mentioned in the texts. The clustering is performed using an Euclidean distance metric.

3.4 Name and face alignment

If a key frame contains a face, then we identify the corresponding ASR or subtitle data that co-occur in a fixed time window with this frame. Further, the names occurring in the textual data are listed as possible names for the frame. As a result, it is possible that an entity mentioned in the text is suggested for several

[1] http://en.wikipedia.org/wiki/Category:British_royal_weddings

[2] http://ariadne.cs.kuleuven.be/TERENCEStoryService/

Table 1: Names and faces alignment results on subtitle vs. ASR data on events

	Subtitle			ASR		
	P	R	F_1	P	R	F_1
Textual	38.095	21.622	27.586	36.585	17.857	24
EM	41.304	25.676	31.667	40.426	22.619	29.008

Table 2: WinDiff score on event identification on subtitle vs. ASR data on the union setting

Subtitle	ASR
11.06	13.80

key frames. However, when there is no corresponding text, or when the text does not contain person entities, no name is suggested for the key frame.

Name and face alignment in royal wedding video is difficult and complicated since the video contains many other faces of persons mentioned above. Sometimes the anchor or designer talks about the couple involved in the wedding, but there is no appearance of this couple in the corresponding video key frame as shown in figure 3.

We minimize this problem of name and face alignment by using the EM algorithm cited in (Pham et al., 2010). Alignment is the process of mapping the faces in the video to the names mentioned in the textual data. For each frame, the most probable alignment scheme has to be chosen from all possible schemes. The EM algorithm has an initialization step followed by the iterative E- and M-steps. The E-step estimates the likelihood of each alignment scheme for a frame, while the M-step updates the probability distribution based on the estimated alignments over all key frames of the video.

3.5 Event identification in subtitle and ASR data with person involvement (if any)

Once the system has learned the events from the Wikipedia data, it identifies the events from the subtitles. The process is as follow: the system scans each subtitle for the key words from the event list. If the key word appears in the subtitle data, then it is treated as the occurrence of the event and stores the set of frames that co-occur with that subtitle. The name and face alignment module already might have yielded a list of names present in this subtitle if there is any person involved. If that is the case, then the names are assigned to the events identified.

The same process is repeated using ASR data.

4 Experimental setup

In this section, we describe the dataset, experimental setup and the metrics used for evaluation.

4.1 Datasets and ground truth annotations

The data used in our experiments is the DVD on Britain's Royal Weddings published by the BBC. The duration of this video is 116 minutes at a frame rate of 25 frames per second, and the frame resolution is 720x576 pixels. Frames are extracted at the rate of one frame per second using the ffmpeg tool (ffmpeg, 2012). Faces in the frames are annotated manually using the Picasa tool for building the ground truth for evaluation. This tool is very handy and user-friendly to tag the faces. We have found that there are 69 persons including British wedding couples in the video. The subtitles came along with the DVD which are already split into segments of around 3 seconds. We use the (FBK, 2013) system to obtain the ASR data of the videos. Since the (FBK, 2013) system takes only sound (.mp3 file) as input, we have converted the video into a mp3 file using (ffmpeg, 2012). The obtained ASR data is then in XML format without any sentence boundaries so we have converted the ASR data into segments in the range of three seconds, which is standard when presenting subtitles in video. It is clear that the ASR transcription contains many words that are incorrectly transcribed. It is also visible that the ASR system does not recognize or misspells many words from the actual speech. As mentioned above, we have built a gazetteer of the couples' names. A set of events are recognized by our system as being important in the context of weddings. To evaluate the quality of these events, the events in the video were annotated by two annotators independently. This annotation includes the actual event, and the start and end times of the event. These two sets with annotations form the groundtruth. To be able to compare the system generated events with the ground truth events, we adopt a two-step approach. First, we combine the corresponding ground truth entries from different annotators into one sequence of frames. Suppose one entry in a ground truth file $(GT(a))$ by one annotator contains the following start (x_a) and end (y_a) time range: $GT(a) : [x_a, y_a]$, and the corresponding entry in the other ground truth file $(GT(b))$ (by the second annotator) contains the following start (x_b) and end (y_b) time range: $GT(b) : [x_b, y_b]$. Merging of the ground truth event ranges can be done in different ways, but we report here on the union of the two ranges.

$$GT(a) \cup GT(b) = [min(x_a, x_b), max(y_a, y_b)] \tag{1}$$

4.2 Evaluation Metrics

Let FL be the final list of name and face alignment retrieved by our system for all the faces detected in all frames, and GL the complete ground truth list. To evaluate the name and face alignment task, we use standard precision (P), recall (R) and F_1 scores for evaluation:

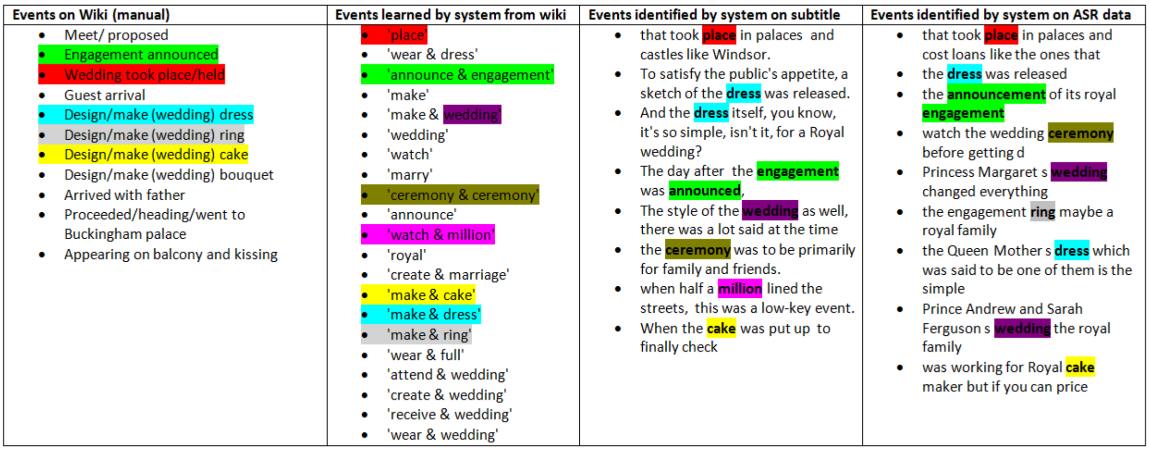
$$P = \frac{|FL \cap GL|}{|FL|} \quad R = \frac{|FL \cap GL|}{|GL|} \quad F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

To evaluate correctness of event segment boundaries, precision and recall are too strict since they penalize boundaries placed very close to the ground truth boundaries. We use the WindowDiff (Pevzner and Hearst, 2002) metric that measures the difference between the ground truth segment GT and the segment SE found by the machine originally designed for text segmentation. For our scenario, this metric is defined as follows:

$$WD(GT, SE) = \frac{1}{M - k} \sum_{i=1}^{M-k} (|b(GT_i, GT_{i+k}) - b(SE_i, SE_{i+k})| > 0) \tag{2}$$

where $M = 7102$, is the number of frames extracted, $k = 1$, is the window size and $b(i, j)$ represents the number of boundaries between frame indices i and j.

5 Results



Events on Wiki (manual)	Events learned by system from wiki	Events identified by system on subtitle	Events identified by system on ASR data
• Meet/ proposed • Engagement announced • Wedding took place/held • Guest arrival • Design/make (wedding) dress • Design/make (wedding) ring • Design/make (wedding) cake • Design/make (wedding) bouquet • Arrived with father • Proceeded/heading/went to Buckingham palace • Appearing on balcony and kissing	• 'place' • 'wear & dress' • 'announce & engagement' • 'make' • 'make & wedding' • 'wedding' • 'watch' • 'marry' • 'ceremony & ceremony' • 'announce' • 'watch & million' • 'royal' • 'create & marriage' • 'make & cake' • 'make & dress' • 'make & ring' • 'wear & full' • 'attend & wedding' • 'create & wedding' • 'receive & wedding' • 'wear & wedding'	• that took **place** in palaces and castles like Windsor. • To satisfy the public's appetite, a sketch of the **dress** was released. • And the **dress** itself, you know, it's so simple, isn't it, for a Royal wedding? • The day after the **engagement** was **announced**, • The style of the **wedding** as well, there was a lot said at the time • the **ceremony** was to be primarily for family and friends. • when half a **million** lined the streets, this was a low-key event. • When the **cake** was put up to finally check	• that took **place** in palaces and cost loans like the ones that • the **dress** was released • the **announcement** of its royal **engagement** • watch the wedding **ceremony** before getting d • Princess Margaret s **wedding** changed everything • the engagement **ring** maybe a royal family • the Queen Mother s **dress** which was said to be one of them is the simple • Prince Andrew and Sarah Ferguson s **wedding** the royal family • was working for Royal **cake** maker but if you can price

Figure 4: Events learned from the Wikipedia data and their identification in the subtitles and ASR by the system

5.1 Evaluation of the extraction of wedding events from Wikipedia

Figure 4 shows which key events typical for royal weddings the system has learned from Wikipedia data and how it found these events in the subtitles and the ASR data. It is seen from figure 4 that the system could not learn many events that are important to wedding events, but the system recognized the events that it has learned quite accurately in the subtitles and ASR data.

5.2 Evaluation of the event segmentation and recognition

Table 2 shows the results of WinDiff score obtained on subtitles versus ASR data on the union setting discussed in 4.1. Though the error rate is more or less the same, it degrades in ASR data which is obviously due to the different ASR errors. The error rate is increased by 2.74% in ASR data using a window size of 1. Here a window size 1 is equivalent to one second so it corresponds to one frame. In this case the system tries to find the event boundaries in each frame and evaluates these against the ground truth event boundaries.

5.3 Evaluation of the name-face alignments

Table 1 shows the result of the name and face alignment given the detected events. Though the result is not quite satisfactory even after applying the EM algorithm, there are many bottlenecks that need to be tackled. Many parts of the video contain interviews. Interviewees and anchors mostly talk about the couples that are married or are to be married, but the couples are not shown which might cause errors in the name and face alignment.

6 Conclusion and future work

In this paper, we have presented ongoing research on event detection in video using ASR and visual data. To some extent the system is able to learn key events from relevant external corpora. The event identification process is quite satisfactory as the system learns from external corpora. If the system would have learnt the events from external corpora good enough, it might identify events very well from subtitle or ASR data. We are interested in improving the learning process from external corpora in further work. Finding event boundaries in the frame sequence corresponding to a subtitle or ASR data where the event is mentioned is still a challenging problem because an event key word might be identified in a subtitle segment or in a sentence which actually may not correspond to what is shown the aligned frames. We have also tried to implement name and face alignment techniques to identify persons involved in the event. As a further improvement of our system, we need to find how to deal with the many interviews in this type of videos which might improve the alignment of names and faces.

References

A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz. 2008. Robust real-time unusual event detection using multiple fixed-location monitors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(3):555–560, March.

James Allan, Stephen Harding, David Fisher, Alvaro Bolivar, Sergio Guzman-Lara, and Peter Amstutz. 2005. Taking topic detection from evaluation to practice. In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04*, HICSS '05, pages 101.1–, Washington, DC, USA. IEEE Computer Society.

James Allan, editor. 2002. *Topic Detection and Tracking: Event-based Information Organization*. Kluwer Academic Publishers, Norwell, MA, USA.

G. Bradski. 2000. Opencv face detector tool. *Dr. Dobb's Journal of Software Tools*. Available at `http://opencv.org/downloads.html`.

Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray. 2004. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22.

FBK. 2013. FBK ASR transcription. Available at `https://hlt-tools.fbk.eu/tosca/publish/ASR/transcribe`.

ffmpeg. 2012. ffmpeg audio/video tool. Available at `http://www.ffmpeg.org`.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of ACL*, pages 363–370.

Brijnesh J. Jain and Klaus Obermayer. 2010. Elkan's k-means algorithm for graphs. In *Proceedings of the 9th Mexican International Conference on Artificial Intelligence: Conference on Advances in Soft Computing: Part II*, MICAI'10, pages 22–32, Berlin, Heidelberg. Springer-Verlag.

Toshio Kawashima, Kouichi Tateyama, Toshimasa Iijima, and Yoshinao Aoki. 1998. Indexing of baseball telecast for content-based video retrieval. In *ICIP (1)*, pages 871–874.

Ken-Hao Liu, Ming-Fang Weng, Chi-Yao Tseng, Yung-Yu Chuang, and Ming-Syan Chen. 2008. Association and temporal rule mining for post-filtering of semantic concept detection in video. *Trans. Multi.*, 10(2):240–251, February.

Qiaozhu Mei and ChengXiang Zhai. 2005. Discovering evolutionary theme patterns from text: An exploration of temporal text mining. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 198–207, New York, NY, USA. ACM.

Lev Pevzner and Marti A. Hearst. 2002. A critique and improvement of an evaluation metric for text segmentation. *Computational Linguistics*, 28(1):19–36.

Phi The Pham, M. F. Moens, and T. Tuytelaars. 2010. Cross-media alignment of names and faces. *IEEE Transactions on Multimedia*, 12(1):13–27, January.

Yong Rui, Anoop Gupta, and Alex Acero. 2000. Automatically extracting highlights for tv baseball programs. In *Proceedings of the Eighth ACM International Conference on Multimedia*, MULTIMEDIA '00, pages 105–115.

Drew D. Saur, Yap-Peng Tan, Sanjeev R. Kulkarni, and Peter J. Ramadge. 1997. Automated analysis and annotation of basketball video. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 176–187.

Cees G. M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 421–430, New York, NY, USA.

Alexander Strehl and Joydeep Ghosh. 2003. Cluster ensembles — a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, March.

Xuanhui Wang, ChengXiang Zhai, Xiao Hu, and Richard Sproat. 2007. Mining correlated bursty topic patterns from coordinated text streams. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, pages 784–793.

Gang Wang, Tat-Seng Chua, and Ming Zhao. 2008. Exploring knowledge of sub-domain in a multi-resolution bootstrapping framework for concept detection in news video. In *Proceedings of the 16th ACM International Conference on Multimedia*, MM '08, pages 249–258, New York, NY, USA. ACM.

Jun Yang, Rong Yan, and Alexander G. Hauptmann. 2007. Cross-domain video concept detection using adaptive SVMs. In *Proceedings of the 15th International Conference on Multimedia*, MULTIMEDIA '07, pages 188–197, New York, NY, USA. ACM.

Dennis Yow, Boon lock Yeo, Minerva Yeung, and Bede Liu. 1995. Analysis and presentation of soccer highlights from digital video. pages 499–503.

Qiankun Zhao, Prasenjit Mitra, and Bi Chen. 2007. Temporal and information flow based event detection from social text streams. In *Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2*, AAAI'07, pages 1501–1506.