Using Sensor Metadata Streams to Identify Topics of Local Events in the City

M-Dyaa Albakour
University of Glasgow, UK
dyaa.albakour@glasgow.ac.uk

Craig Macdonald
University of Glasgow, UK
craig.macdonald@glasgow.ac.uk

Iadh Ounis
University of Glasgow, UK
iadh.ounis@glasgow.ac.uk

ABSTRACT

In this paper, we study the emerging Information Retrieval (IR) task of local event retrieval using sensor metadata streams. Sensor metadata streams include information such as the crowd density from video processing, audio classifications, and social media activity. We propose to use these metadata streams to identify the topics of local events within a city, where each event topic corresponds to a set of terms representing a type of events such as a concert or a protest. We develop a supervised approach that is capable of mapping sensor metadata observations to an event topic. In addition to using a variety of sensor metadata observations about the current status of the environment as learning features, our approach incorporates additional background features to model cyclic event patterns. Through experimentation with data collected from two locations in a major Spanish city, we show that our approach markedly outperforms an alternative baseline. We also show that modelling background information improves event topic identification.

1. INTRODUCTION

Local search is increasingly attracting more demand, whereby the users are interested to find out about places or events in their local vicinity [11]. Local event retrieval is an example of local search where users can retrieve a ranked list of local events of interest, such as music concerts, entertainment events or even protests. Recent work has addressed local event retrieval by using social media activity as a sensor to detect and rank events [1, 13]. However, social media may only cover very popular events as users may not necessarily comment on all events taking place in the city. Therefore, physical sensors that record observations about the status of the environment can provide additional evidence about the events taking place in the city. These sensors can take the form of visual sensors such as CCTV cameras, acoustic sensors such as microphones or possibly environmental sensors.

There is a wealth of research on identifying low-level human activities from acoustic and visual sensors. Often, this involves sensor signal processing to extract sensor features for modelling human activities. For example, Atrey *et al.* [2] developed a Gaussian Mixture Model using a variety of features derived from audio signal processing to classify human activities into vocal classes, such as talking and shouting, and non-vocal classes, such as walking and running. Similar approaches also used audio signal features to identify low-level human activities that are related to security incidents, such as breaking glass or explosions [7]. In addition to using acoustic sensors, several studies have been conducted to identify low-level human activities from videos. Since its introduction in 2002, the TRECVID evaluation campaign [9] has tackled a variety of content-based retrieval tasks from video recordings to support video search and navigation. This includes the semantic indexing of video segments, whereby videos are mapped to concepts, which can be certain objects or human activities [9]. Another related task is multimedia event detection, where the aim is to identify predefined classes of events in the videos. In this task, the existing effective approaches employ classifiers trained with motion features from the videos [10]. Moreover, classifying human interactions identified in video recordings has been studied to detect surveillance-related incidents [5].

Although the aforementioned approaches derive useful semantics about the multimedia content, they only consider low-level human activities. In other words, they provide sensor metadata describing low-level human activities in the physical environment. However, to the best of our knowledge, no previous work has investigated combining these sensor metadata to detect and retrieve higher level complex events taking place in the city, such as music concerts or entertainment shows, which may involve several lower level human actions. In this paper, we propose an approach for combining sensor metadata streams to support local event retrieval. Our major contribution in this paper is devising a supervised machine learning approach that combines sensor metadata to identify the topic of a potential event happening at a particular time in a certain location of the city. The topic corresponds to a set of terms representing a type of events, such as a concert or a protest. Our approach uses features from acoustic, visual and social sensor metadata. We also incorporate background features from past observations to model events that exhibit cyclic patterns such as traffic jams at peak times. To develop our supervised approach, we perform two necessary steps. First, we obtain event annotations on a pool of candidate video and audio recordings of two vibrant locations in the centre of a major Spanish city over a period of two weeks. Second, we use the obtained annotations to map typical events taking place in those locations into coherent topics using a topic mod-

SIGIR '15, August 09 - 13, 2015, Santiago, Chile
 ISBN 978-1-4503-3621-5/15/08...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767837 .

elling technique. Through experimentation with the data collected, we evaluate the accuracy of our event topic identification approach and shows that it markedly outperforms an alternative baseline. The results also show the effectiveness of the background features in improving the accuracy of event topic identification.

2. PROBLEM STATEMENT

In this paper, we tackle the problem of event topic identification. The aim is to combine the sensor metadata observations captured at different locations in a city to identify topics of potential high-level events taking place within certain locations. Formally, for a location l_i in a city, we denote the sensor metadata observations captured at time t_j in that location l_i by the vector $\overrightarrow{N}_{\langle l_i,t_j\rangle}$. The sensor metadata observations may include the crowd density identified from captured videos in the location, low-level audio events identified from the acoustic sensors installed in the location or social media activities, such as tweets posted by people at the location. The problem of event topic identification is to use the vector $\overrightarrow{N}_{\langle l_i,t_j\rangle}$ to map the tuple of time and location $\langle l_i, t_j\rangle$ to a certain topic $p_x \in \mathcal{P}$ described by a set of terms T_x; where $\mathcal{P}$ is a set of predefined topics.

In a previous work [1], the textual content of public tweets has been used as the only source of sensor metadata observation to identify topics of local events. Although this has worked well on popular events that attract social media activities, it did not work as well on more localised events that may not attract coverage on the social media. To alleviate this shortcoming, we introduce *physical* sensor metadata streams that can provide an additional evidence for the topic of an event, namely video and audio metadata observations. However, this requires understanding the semantics of visual scenes or audio recordings, which remains an open challenge. Indeed, there is no known taxonomy that maps sensor metadata to topics of high level events. To address this challenge, we propose to learn the topic associated with a tuple $\langle l_i, t_j\rangle$ from *labelled* training data using features extracted from the sensor observations $\overrightarrow{N}_{\langle l_i,t_j\rangle}$.

For this purpose, and to collect labelled training data, we obtain event annotations on a pool of videos that are identified as potential candidates to contain events. Furthermore, to extract a predefined set of coherent event topics, we apply topic modelling on the descriptions of the annotated events. We detail the event annotation and the topic extraction in Section 4. Next, we describe the sensor data collection.

3. SENSOR DATA COLLECTION

Our study considers two locations in the city centre of Santander in Spain. The first location is the geographical and business heart of the city; it is a major square opposite to the municipality building. The second location is a popular open market in the city, where hundreds of people go every day for shopping, located behind the municipality building. Both locations represent vibrant and busy areas, where we expect to observe high-level events of interest such as music concerts, entertainment shows or even protests. The data collection started since the 11^{th} October 2013 in both locations.

Table 1 provides a summary of the sensor data collection and the metadata produced by processing the output from the microphones and the camera in each location. For producing the audio metadata, a supervised classifier using feed forward multilayer perceptron network and low-level audio features, such as those described in [6], was developed for each of the following 6 audio classes described in Table 1: "crowd", "traffic, "music", "applause", "speaker", and "siren". For video metadata, the video was processed for crowd analysis where we calculate the crowd density, in desired areas, by estimating the foreground components of the video. In addition to the acoustic and visual sensors, we collected parallel social media activity in the city. In particular, using the Twitter Public Streaming API[1], we obtained tweets related to each location (as identified by their geo-locations).

Table 1: Summary of sensor data collection

Locations	2 (square & market)
Physical sensors	A camera and microphones (in each location)
Raw output	1600x1200 video @ 20fps 16Khz audio @ 64kbits/s (audio is multiplexed with the video)
Audio metadata	classification scores for 6 audio classes (i) "crowd": noise from a crowd of people (ii) "traffic": car and road noises (iii) "music": music played outdoors (iv) "applause": applause, yelling or cheering (v) "speaker": speech over loud speakers (vi) "siren": noise of police cars & ambulances
Video metadata	crowd density in the scene
Twitter	geo-tagged tweets within each location

4. EVENT POOLING AND ANNOTATION

In this section, we describe our approach for obtaining event annotations on the recordings collected from the two locations. Recall from Section 2, that our ranking units (the documents) are tuples of time and location. Each tuple represents a segment of recordings at a location. The length of the segment, the sampling rate to obtain the tuples, can be predefined and we follow [1] in setting the sampling rate to 15 minutes. Coarser- or finer-grained sampling rate can be investigated in future work for different types of events e.g. emergency events may require a finer-grained sampling.

For annotation, we consider a period of 2 weeks starting from 19 October 2013, around a week after the start of the data collection (11 October 2013) to allow the estimation of background features. Since it is expensive to examine all recordings and annotate them with events, we employ a pooling approach [4], as commonly used in IR campaigns, such as TREC.

For pooling, we identify candidate segments of videos where high-level events may have occurred by applying the change component of the event retrieval framework in [1]. In particular, the change component of this framework identifies segments where sensor metadata observations change unusually in a location, e.g. unusual change in crowd density. We use 4 different types of sensor metadata observations to generate the pool (a subset of those listed in Table 1): (i) the median values of the video crowd density, (ii) the median values of the crowd audio classification score, (iii) the median values of the music audio classification score, and (iv) the total number of tweets posted. As a result, we obtain a total of 155 candidate segments. The video recording software produced videos with lengths of either 30 minutes or 1 hour, and the total number of video recordings that correspond to the 155 segments are 69 videos.

The generated candidate segments of videos were then annotated by two groups of human annotators, English and Spanish annotators, who were asked to examine the videos,

[1] https://dev.twitter.com/streaming/public

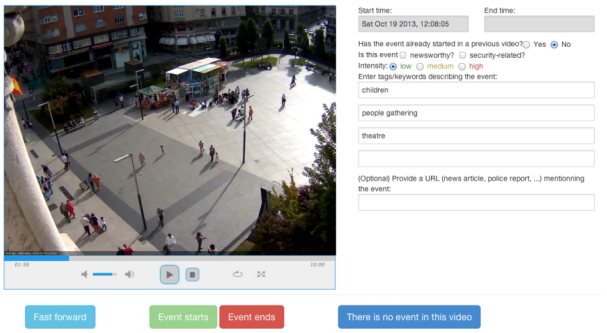

Figure 1: A snapshot from the annotation interface.

Table 2: Statistics of annotated videos

Annotators	unique videos	annot. ratio	mutliple annotations	Agreement
4 English	29	29/69= 42%	1 video	100%
5 Spanish	47	47/69= 68%	13 videos	77%
Both	55	55/69= 79%	21 videos	71%

describe events that they observe by typing in terms, and rate their intensity on a 3-point scale (Low, Medium, and High) according to how likely they are to generate public interest. The intensity is akin to graded relevance used in traditional IR evaluation approaches [12]. We provided the annotators with a web-based interface, of which we show a snapshot in Figure 1.

Statistics for the obtained annotations are summarised in Table 2. From the last row we observe that we obtain a total of 55 annotated videos, of which 21 were annotated by more than one annotator. The agreement between annotators is estimated by converting the intensity levels to binary decisions, using "Medium" as a threshold. We observe that a reasonable agreement is achieved in all cases (lowest is 71%), which gives us confidence in the annotations obtained.

For the set of annotated pooled segments, we obtain terms describing events that were identified in these segments. For each annotated segment, we construct a virtual document that consists of all of the terms provided by the annotators. Since the pooled videos were annotated by both spanish and english annotators, these virtual documents are bilingual and contain English and/or Spanish terms. To cluster events into various topics, we propose to use topic modelling on the document collection of all constructed virtual documents of terms. We use the Latent Dirichlet Allocation (LDA) topic modelling implemented in the Mallet toolkit [8]. In Table 3, we list the top terms of 7 identified topics from the English annotations only. From the table, we can observe that the identified topics are reasonable where we see some interesting associations of terms that describe typical high-level events taking place in the square and the market, e.g. 'demonstration' and 'show' in topic 4, and 'children' and 'entertainment' in topic 6.

5. LEARNING EVENT TOPICS

In this section, we discuss our supervised approach for event topic identification, where the aim is to identify the topic of a segment $\langle l_i, t_j \rangle$ using the sensor metadata observations $\overrightarrow{N}_{\langle l_i,t_j \rangle}$. To train our supervised approach, we construct a *labelled* dataset of event topics from the annotated video pool collected in Section 4. The labelled data consists of segments (tuples of time and location) labelled with either an event topic or with the label 'no event' indicating that no event of interest has occurred in the corresponding time and location. We labelled each annotated segment in the pool to the most probable topic according to the LDA topic modelling configured by setting the number of topics to 7.[2] Unlabelled segments or where the annotators did not identify any event are associated to the 'no event' label. To illustrate the volume of the data and the distribution of labels, we detail in Table 4 the number of segments for each label when using topic modelling on all Spanish and English annotations and setting the number of topics to 7.

Table 3: Topics identified with topic modelling using the English annotations

Topic	Top terms of the topic
Topic 1	loudspeaker people fanfare police drums procession
Topic 2	microphone rings speech public claps
Topic 3	gathering plaza people booth theatre music
Topic 4	demonstration sitting event sound speak show
Topic 5	market protest cars ongoing children fair
Topic 6	children people shopping middle entertainment
Topic 7	music singing playing guy bells whistles

Table 4: Distribution of labels

Lab.	#	Lab.	#	Lab.	#	Lab.	#
top.1	12	top.3	2	top.5	0	top.7	11
top.2	8	top.4	32	top.6	24	no event	66

We consider the problem of identifying the topic of a pooled segment as a classification task. Using the constructed labelled data, we train a binary classifier for each of the labels with features derived from various sensor metadata streams. Our intuition is that such labelled data would allow us to learn the semantics of a combination of sensor metadata. In other words we aim to match sensor metadata to topics defined using the annotations. For training the classifier, we investigate two main sets of features for the segment, *observation features* and *background features*. Table 5 summarises those features. The *observation features* are extracted from the sensor metadata observed in the location and time corresponding to the segment. The *background features* aim to model past observations and cyclic patterns of activities that take place over time in the same location. The intuition is that some events are periodic and exhibit a long-term pattern, e.g. traffic jams at peak times resulting in a high traffic audio classification score, or entertainment shows taking place in the square at the same time on the weekends. Modelling cyclic patterns, i.e. daily and weekly cycles, from the sensor metadata observations would enable the supervised classifier to identify recurring background events or noise which are not of interest such as traffic jams. Similarly, it would help to identify recurring events of interest such as entertainment shows.

Using the labelled dataset of segments along with the features described in Table 5, we apply supervised machine learning to learn a binary classifier for each label. In particular, we experiment with Random Forests [3] as a learning algorithm.[3] Next we conduct a number of experiments to evaluate the accuracy of our classifier and the effectiveness of the various devised features.

6. EXPERIMENTS

To evaluate our approach for identifying the topic of a candidate segment, we use the dataset of labelled segments

[2]We use 7 topics since we have observed that with this setting we obtain the most coherent topics after experimenting with other alternatives (varying the number of topics between 5 and 10)

[3]We also experimented with other supervised machine learning algorithms, such as naive Bayes and SVM, but due to the space limit, we only report results with Random Forests since they achieve the best performances and the conclusions with other algorithms are similar.

Table 5: Features devised for topic identification

8 Observation features		
Audio features	6	median of the classification score for each audio class (crowd, traffic, music, applause, speaker, siren)
Video features	1	median of the crowd density score
Twitter features	1	number of tweets geotagged within the location in the past one hour
16 Background features		
Daily aggregates	8	for each of our 8 observation features its daily median from all available past observations at the same time from previous days
Weekly aggregates	8	for each of our 8 observation features its median from all available past observations at the same time on the same day of previous weeks
Total	24	

Table 6: Performance of topic identification

Approach	F_1 Accuracy	Precision	Recall
Majority baseline	0.254	0.181	0.426
Obs. Feat.	0.686	0.705	0.697
Obs. & Daily	0.740	0.759	0.761
Obs. & Weekly	0.715	0.715	0.729
Obs. & All background	**0.766**	**0.781**	**0.762**

described in Section 5. We perform a 10-fold cross validation and report the average accuracy across all labels (a label for each topic and the label 'no event'). In addition to using different instantiations of our classifier, we also compare our classifier to an alternative baseline. The "majority" baseline assigns the most common label in the training data to the segments in the testing data. Table 6 summarises the results.

We observe from the table that all instantiations of our approach are markedly better than the majority baseline. In particular, when using only the observation features our approach achieves an F_1 accuracy of 0.686. We also observe that this performance further increases when using the background features. Indeed the best performance is achieved when using all background features along with the observation features ($F_1 = 0.766$). This illustrates that modelling cyclic patterns by aggregating sensor metadata from previous observations helps in better identifying whether a candidate segment represents an event and in identifying the topic of an event.

Furthermore, we conduct an ablation study to identify which features are more effective for topic identification. We remove one of our 8 observation features when learning the classifier. We report the results in Table 7. For example, the row headed "- (Audio crowd)" means that we use all the observation features apart from the audio crowd score. We observe that removing any of the features results in a degrading of performance for accuracy and precision. This is an interesting observation and highlights the importance of having rich metadata describing the environment for identifying the topics high-level events. However, we also observe that the performance degrades most when removing the audio crowd score and the crowd density features. This suggests that the crowd level, as detected by the acoustic or visual sensors, is important to identify events and to distinguish their topic.

7. CONCLUSIONS

In this paper, we proposed an approach for combining sensor metadata streams to identify the topics of events happening within a city. Our approach trains a classifier to identify topics of candidate segments of recordings. Our results are promising and show that combining features from a variety of sensors (acoustic, visual and social) and modelling cyclic patterns from past observations provides the best accuracy for event topic identification. These results pave the way towards more robust implementations of local event retrieval that harness both physical and social sensor streams.

Table 7: Results of the ablation study

Model	F_1 Accuracy	Precision	Recall
All observation features	**0.686**	**0.705**	**0.697**
- (Audio crowd)	0.635	0.624	0.635
- (Audio traffic)	0.681	0.678	0.691
- (Audio applause)	0.680	0.678	**0.697**
- (Audio music)	0.685	0.682	**0.697**
- (Audio speaker)	0.657	0.656	0.665
- (Audio siren)	0.656	0.655	0.665
- (Video crowd)	0.652	0.651	0.665
- Twitter	0.682	0.677	0.697

Acknowledgments

This work has been carried out in the scope of the EC cofunded project SMART (FP7-287583) and also in the integrated Multimedia City Data project funded by the Economic and Social Research Council.

8. REFERENCES

[1] M.-D. Albakour, C. Macdonald, and I. Ounis. Identifying local events by using microblogs as social sensors. In *Proc. of OAIR'13*.

[2] P. K. Atrey, M. Maddage, and M. S. Kankanhalli. Audio based event detection for multimedia surveillance. In *Proc. of ICASSP'06*.

[3] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[4] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. of SIGIR'04*.

[5] F. Chen and W. Wang. Activity recognition through multi-scale dynamic bayesian network. In *Proc. of VSMM'10*.

[6] J. Dennis, Q. Yu, H. Tang, H. D. Tran, and H. Li. Temporal coding of local spectrogram features for robust sound recognition. In *Proc. of ICASSP'13*.

[7] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini. Automatic sound detection and recognition for noisy environment. In *Proc. of EUSIPCO'00*.

[8] A. K. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[9] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quénot. TRECVID 2014 – An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proc. of TRECVID'14*.

[10] S. Phan, T. D. Ngo, V. Lam, S. Tran, D.-D. Le, D. A. Duong, and S. Satoh. Multimedia event detection using segment-based approach for motion feature. *Journal of Signal Processing Systems*, 74(1):19–31, 2014.

[11] G. Sterling. Study: 43 percent of total google search queries are local. http://searchengineland.com/study-43-percent-of-total-google-search-queries-have-local-intent-135428. Accessed: 5 October 2012.

[12] E. M. Voorhees. Evaluation by highly relevant documents. In *Proc. of SIGIR'01*, pages 74–82, 2001.

[13] M. Walther and M. Kaisser. Geo-spatial event detection in the twitter stream. In *Advances in Information Retrieval*, pages 356–367. Springer, 2013.