Concept Language Models and Event-based Concept Number Selection for Zero-example Event Detection

Damianos Galanopoulos
Information Technologies Institute (ITI), CERTH
Thermi, Greece, 57001
dgalanop@iti.gr

Foteini Markatopoulou
Information Technologies Institute (ITI), CERTH
Thermi, Greece, 57001
Queen Mary University of London
markatopoulou@iti.gr

Vasileios Mezaris
Information Technologies Institute (ITI), CERTH
Thermi, Greece, 57001
bmezaris@iti.gr

Ioannis Patras
Queen Mary University of London
Mile end Campus, UK, E14NS
i.patras@qmul.ac.uk

ABSTRACT

Zero-example event detection is a problem where, given an event query as input but no example videos for training a detector, the system retrieves the most closely related videos. In this paper we present a fully-automatic zero-example event detection method that is based on translating the event description to a predefined set of concepts for which previously trained visual concept detectors are available. We adopt the use of Concept Language Models (CLMs), which is a method of augmenting semantic concept definition, and we propose a new concept-selection method for deciding on the appropriate number of the concepts needed to describe an event query. The proposed system achieves state-of-the-art performance in automatic zero-example event detection.

CCS CONCEPTS

• **Information systems** → **Query representation**; **Video search**;

KEYWORDS

Zero-example multimedia event detection; Video search; Query representation

ACM Reference format:
Damianos Galanopoulos, Foteini Markatopoulou, Vasileios Mezaris, and Ioannis Patras. 2017. Concept Language Models and Event-based Concept Number Selection for Zero-example Event Detection. In *Proceedings of ICMR '17, Bucharest, Romania, June 06-09, 2017,* 5 pages.
https://doi.org/10.1145/3078971.3079043

1 INTRODUCTION

Multimedia-event detection is a very important task that deals with automatically detecting the main event presented in a video. As a video event we consider a complex activity involving people interacting with other people and/or objects, e.g.,"Renovating a home". Typically, multi-class classification is used to train event detectors on ground-truth annotated video samples. However, collecting ground-truth annotated data is difficult and time consuming. As a result, the more practically applicable but also more challenging zero-example event detection task has gained significant attention. The objective of this task is to retrieve the most closely related videos from a large video collection, given any abstract event description for which training samples are not available.

Recent studies typically start by analysing the textual event description so as to transform it to a meaningful set of keywords. At the same time, a predefined set of concepts is used, in the one hand to find which of these concepts are related to the extracted keywords and consequently to the event description, and on the other hand, to train visual concept detectors that will be used to annotate the videos with these semantic concepts. The distance between the event's concept vector and each videos concept vector is calculated and the videos with the smallest distance are selected as being the most closely related to the given event. In this work we improve such a typical system in the following ways: i) We adopt an efficient way for augmenting the definition of each semantic concept in the concept pool, ii) We present a new strategy for deciding on the appropriate number of concepts for representing the event query, iii) We combine these in a zero-example event detection method that outperforms the state-of-the-art techniques.

2 RELATED WORK

Zero-example event detection is an active topic with many literature works proposing ways to build event detectors without any training samples using solely the event's textual description. Research towards this problem was mainly triggered a few years ago when the TRECVID benchmark activity introduced the 0Ex task as a subtask of the Media Event Detection (MED) task [9]. A similar to zero-example event detection problem, known as zero-shot learning (ZSL), also appears in the image recognition task. A new unseen category, for which training data is not available, is asked to be detected in images [3, 7, 12]. It should be noted that although the two problems have many common properties, zero-example event detection is a more challenging problem as it focuses in more

ICMR '17, June 06-09, 2017, Bucharest, Romania

ACM ISBN 978-1-4503-4701-3/17/06…$15.00
https://doi.org/10.1145/3078971.3079043

complex queries, where multiple actions, objects and persons interact with each other compared to the simple object or animal classes that appear in ZSL [19].

The problem of zero-example event detection is typically addressed by transforming both the event textual description and the available videos into concept-based representations. Specifically, a large pool of concept detectors is used to annotate the videos with semantic concepts, the resulted vectors, a.k.a. model vectors, contain the scores indicating the degree that each of the concepts is related to the video. The query description is analysed and the most related concepts from the pool are selected. Finally, the distance between the model vectors and the event concept vectors is calculated and the most related videos are retrieved [1, 4, 18, 21].

Concept detectors are typically trained on external ground-truth annotated datasets using for example deep nets (DCNNs) or low-lever features from different modalities [10]. The simpler way of translating an event query into keywords is space separating the event's textual description, removing the stop-words and using simple NLP rules [16]. Then, each of the keywords is compared with each of the concepts and the top-related concepts are selected to represent the event. Typically, a fixed number is used to decide how many concepts will be selected for each event query [2, 18]. However, adjusting the number of concepts based on the textual description has not been investigated. A semi-automatic approach is proposed by [2]; initially, the system automatically detects the concepts that are related to an event query. Subsequently, a human manually removes the noisy ones. The authors argue on the importance of such human intervention due to the big influence that the selection of the wrong concepts for an input query has on the system's accuracy. Furthermore, comparing each keyword with a single concept may be suboptimal. In some works the augmentation of concepts with synonyms is proposed, while the authors in [18] proposed a method where Concept Language Models are built using online sources, such as Google and Wikipedia, for augmenting the concept definitions with more information. In [20], logical operators are used to discover different types of composite concepts, which leads to better event detection performance. More clever ways of augmenting the concept pool should be found. In [4] instead of calculating concept-related event and video vectors both the videos and the event queries are embedded into a distributional semantic space, then the similarity between these representations is measured. Xiaojun et al. [11] proposed a zero-example event detection method, which initially learns a skip-gram model in order to find the semantic correlation between the event description and the vocabulary of concepts. Then, external video resources are retrieved and dynamic composition is used in order to calculate the optimal concept weights that will be aligned with each testing video based on the target event. Although this approach presents very promising results, the retrieval of external videos and the concept weight calculation are computational expensive. Finally, works in [14] and [13] focus on the improvement of the system's retrieval accuracy by using pseudo-relevance feedback.

3 PROPOSED APPROACH

In this section we present a fully automatic zero-example event detection system as presented in Fig. 1. The proposed system takes as input the event kit, i.e., a textual description of the event query, and retrieves the most related videos from the available event collection. An example of an event kit is presented in Fig. 2. As it can be seen it is a textual description of the requested event that includes the event's title, a short definition of the event and visual and audio cues that are expected to appear in those videos that contain this event. The complete procedure is split into two major components. The first component (upper part of Fig. 1) builds the event detector i.e., a vector of the mostly related concepts based on the event kit. The second component (lower part of Fig. 1) calculates the video model vectors, i.e., annotates the videos with semantic concepts. Finally, the output of the two components is compared, using a similarity measure, and the video model vectors that are closer to the concept event detector are retrieved.

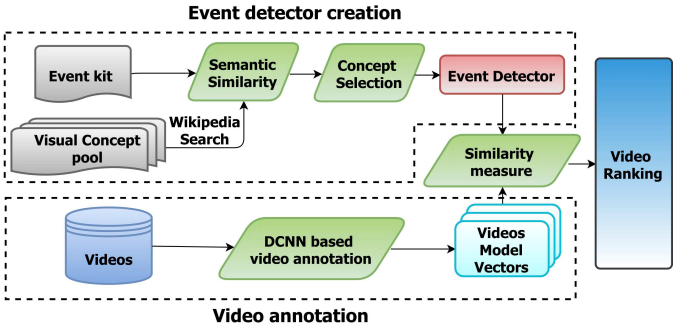

Figure 1: The proposed pipeline for zero-example event detection.

3.1 Building an Event Detector

An event detector is a k-element vector $\mathbf{d}$ of the most related concepts to the event query. Each element indicates the degree that each of the k concepts is related to the target event query. To calculate the event detector we propose a method as follows (algorithm 1): Firstly, we check if the entire event title is semantically close to any of the available concepts from the concept pool, i.e., we check if the semantic relatedness between the event title and each of the concepts is above a (rather high) threshold. If so, we consider that the event is well-described entirely by this (or those) concepts and the relatedness value of these is used to form the event detector $\mathbf{d}$. The Explicit Semantic Analysis (ESA) measure [17] is used to calculate the semantic relatedness of two words or phrases. The ESA measure calculates the similarity distance between two terms by computing the cosine similarity between their corresponding weighted vectors of Wikipedia articles. We choose this measure because it is capable of handling more complex phrases and not only simple words. If the above process does not detect any related concepts, then an Event Language Model (ELM) and Concept Language Models (CLM) are built as follows.

(a) Event Language Model (ELM) and Concept Language Model (CLM). An ELM is a set of N word and phrases that are extracted from the event kit. We build an ELM using the event title, and the visual and audio cues, by simply space separating them. A CLM is a set of M words or phrases that are extracted w.r.t. to a specific concept definition. A CLM is built for each concept using

Event Title: Attempting a bike trick.

Definition: One or more people attempt to do a trick on a bicycle, motorcycle, or other type of motorized bike. To count as a bike for purposes of this event, the vehicle must have two wheels (excludes unicycles, ATVs, etc).

Visual cues:
- **Scene**: outdoors, often in skate park, parking lot or street.
- **Objects/People**: person riding a bike, bike, ramps, helmet, concrete floor, audience.
- **Activities**: riding bike on one wheel, standing on top of bike, jumping with the bike (especially over or onto objects), spinning or flipping bike.

Audio cues:
- **Audio**: sounds of bike hitting surface during the trick, audience cheering.

Figure 2: The event kit text for the event class "Attempting a bike trick"

Algorithm 1: Building an event detector

```
Input  : Event_Title; List_of_visual_concepts
Output : Event detector d
i ← 0;
forall list_of_visual_concepts do
    score ← ESA(visual_concept_Title, Event_Title)
    if score > Threshold then
        d(i) ← score;
        i++;
if size of D > 0 then
    return d ;
else
    build ELM;
    forall list_of_visual_concepts do
        build CLM;
        W ← ESA(ELM,CLM);
        vector_of_concepts_scores(j) ← D_H(W) ;
    S ← sort(vector_of_concepts_scores);
    area ← calculate the area below S;
    for i ← 2 to sizeof (S) do
        tmp ← calculate the area below the curve between S_1, S_i;
        if tmp > area*X% then
            break;
    d = S_1, ..., S_i;
    return d ;
```

the top articles in Wikipedia. The retrieved articles are transformed in a Bag-of-Words (BoW) representation from which the top-M words, which are the most characteristic words of this particular visual concept, are kept. For example, the top retrieved words for the concept "*palace*" are "*palace*","*crystal*", "*theatre*", "*season*", "*west*", "*east*", "*spanish*", "*gates*", "*hotel*". After building the ELM and CLMs, we calculate a single value per concept that denotes the semantic relation of this concept with the ELM. To do that, for each CLM we calculate a $N \times M$ distance matrix W. Each element of the matrix contains the semantic relatedness (in terms of the ESA measure) between pairs of words appearing in the ELM and CLM. Given the matrix W a single score is calculated by applying to W the Hausdorff distance, defined as $D_{\mathcal{H}}(\text{EML}, \text{CLM}) = \text{median}\left(\max_{1 \le j \le N}(\mathbf{d_j})\right)$ where $\mathbf{d_j} = [W_{1,j}, W_{2,j}, \ldots, W_{M,j}]$. The single values calculated per concept, by repeating the above process for every CLM, are concatenated into a single k'-element vector $\mathbf{d'}$ and a process is followed for deciding the appropriate number of concepts that will be finally kept for representing the event query.

(b) Event-based concept number selection. In contrast to [18] and [2], where the number of selected concepts is fixed across the different events and motivated by statistical methods such as PCA [15], where a fraction of components are enough to efficiently or even better describe the data, we propose a statistical strategy that decides on the appropriate number of concepts k, where $k \le k'$, that should be kept for an event query. Specifically, our strategy orders the vector of concepts scores $\mathbf{d'}$ in descending order, constructs an exponential curve, and then selects the first k concepts so that the corresponding area under the curve is at the $X\%$ of the total area under the curve. This procedure, consequently returns different number of selected concepts for different target events. For example for the event "*Attempting ordering the a bike trick*" the selected concepts are the following four: "*ride a dirt bike*", "*mountain biking*", "*put on a bicycle chain*", "*ride a bicycle*", while for the event "*Cleaning an appliance*" only the concept "*clean appliance*" is selected. The final event detector is a k-element vector that contains the relatedness scores of the selected concepts.

3.2 Video Annotation and Retrieval

Initially, each video is decoded into as set of keyframes at fixed temporal intervals. Then, a set of pre-trained concept-based DCNNs are applied to every keyframe and each keyframe is represented by the direct output of those networks. Finally, a video model vector is computed by averaging (in terms of arithmetic mean) the corresponding keyframe-level representations. Each element of a model vector indicates the degree that each of the predefined concepts appears in the video.

The distance between an event detector and each of the video-level model vectors is calculated, and the h videos with the smallest distance are retrieved. As distance measure we choose the histogram intersection, which calculates the similarity of two discretized probability distributions and is defined as follows:
$K_{\cap}(a, b) = \sum_{i=1}^{k} min(a_i, b_i)$.

4 EXPERIMENTAL RESULTS

We use the TRECVID MED14TEST dataset [9] that contains approximately 25.000 videos. We evaluate all the methods on the 20 MED2016 [5] Pre-Specified events (E021-E040) for which event kits are provided. We use a concept pool that consist of 13.488 semantic concepts collected from two different sets: i) 12.988 concepts from the *ImageNet* "fall" 2011 dataset [8] and ii) 500 high level concepts from the *EventNet* [6] dataset. Each video was decoded into 2 keyframes per second and each keyframe was annotated with all the above concepts. In order to obtain scores regarding the 12.988 ImageNet concepts we use the pre-trained GoogLeNet provided by [22], while we use the EventNet [6] network for gathering scores w.r.t. the 500 event-based concepts. The final video model vector is calculated by averaging the scores of the keyframe representations in terms of arithmetic mean. We evaluate all the methods in terms of the Mean Average Precision (MAP).

In our first set of experiments we investigate how different parameters of our method affect the final performance. Firstly, in Table 1, we compare 3 different types of ELMs (CLM was kept the same for all of the three ELMs). The first ELM is built solely from the event title, the second ELM uses both the event title and visual cues

ELM type	Event Title	Visual	Audio-Visual
MAP	0.091	0.122	**0.133**

Table 1: Comparison between different types of ELM

CLM type	Concept Title	Wikipedia
MAP	0.092	**0.133**

Table 2: Comparison between different types of CLM

and the third one uses the event title, visual and audio-visual cues. According to Table 1 the more information is given for building an ELM the better the overall accuracy, i.e., the third ELM that uses the complete event kit description (except for the event definition) outperforms the other two that use sub-parts of the event kit.

Similarly, in Table 2 and Figure 3 we compare 2 different types of CLMs. The first CLM uses solely the concept name, along with any available description of it. The second CLM augments the concept name with terms in Wikipedia as described in Section 3.1; the top-10 words of the BoWs representation from the top-10 retrieved documents are used. Similar to the ELMs, the more information provided for building a CLM the better the overall accuracy, i.e., augmenting a concept with information captured from Wikipedia improves the video detection performance. We noticed that 7 out of 20 events had the same performance irrespective of the used CLM types. This happened because existing concepts in our initial pool can describe adequately these specific events.

Figures 4 and 5 present how the different quota of the area under the curve (AUC) (which implies different number of the

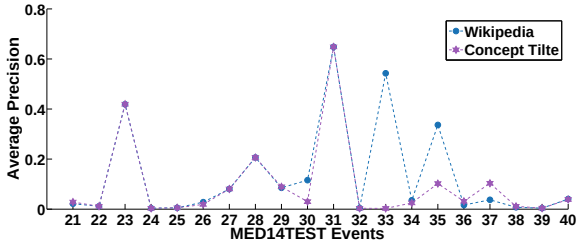

Figure 3: The performance for different types of CLM for the 20 MED14TEST events

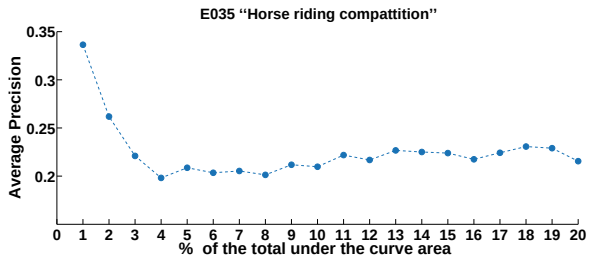

Figure 4: The performance of different quota of under the curve area for the E035 event

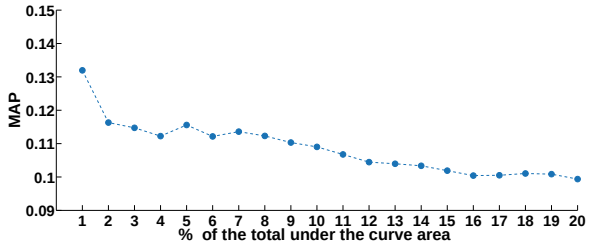

Figure 5: The overall performance of different quota of under the curve area for MED14TEST dataset

Method	MAP
AutoSQGSys [13]	0.115
Concept Bank [1]	0.129
Tzelepis et al. [18]	0.119
Proposed System	**0.133**

Table 3: Comparison between different zero-example event detection systems

top-k selected concepts in every event), affects the performance of our method. In Fig. 4 we observe that the better AP w.r.t the event "Horse riding competition" are achieved for small values of the AUC. This indicates that selecting more concepts that are not highly related with the event query adds noise to the retrieval process that consequently reduces the overall accuracy. Similar conclusions for the overall performance can be reached w.r.t. Fig. 5. The best performance is achieved when the 1% of the AUC is chosen. In this case the average number of selected concepts is 20.1, but each event has different number of concepts. For example the event "*Attempting a bike trick*" needs only 4 concepts while event "*Winning a race without a vehicle*" needs 38 concepts.

In our second set of experiments, we compare the proposed method with the following three state-of-the-art ones: i) AutoSQGSys System [13], ii) Concept Bank system [1] and iii) Tzelepis et al. zero-example method [18], where a fixed number of selected concepts was used. The results of [13] and [1] are picked up from the corresponding papers while the [18] method was re-implemented in order to be suitable for our experiment set-up. According to Table 3 the proposed method outperforms all of the other approaches reaching a MAP of 0.133.

5 CONCLUSION

In this paper we present a fully-automatic method for zero-example video event detection. The augmentation of the concept descriptions with extra information in combination with the proposed strategy for deciding on the appropriate number of concepts for representing the event query outperforms all the state-of-the-art approaches presented in this paper.

ACKNOWLEDGMENTS

This work was supported by the EU's Horizon 2020 research and innovation programme under grant agreements H2020-693092 MOVING and H2020-687786 InVID.

REFERENCES

[1] Y.J Lu, H. Zhang, M. de Boer, C.W. Ngo. 2016. Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval (ICMR '16)*. ACM, New York, NY, USA, 127–134.

[2] Y.J. Lu, H. Zhang, M. de Boer, C.W. Ngo. 2016. Event detection with zero example: select the right and suppress the wrong concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 127–134.

[3] M. Elhoseiny, B. Saleh, A. Elgammal. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *Computer Vision (ICCV), IEEE Int. Conf. on*. IEEE, 2584–2591.

[4] M. Elhoseiny, J. Liu, H. Cheng, H. Sawhney, A. Elgammal. 2015. Zero-shot Event Detection by multimodal distributional semantic embedding of videos. *arXiv preprint arXiv:1512.00818* (2015).

[5] G. Awad, J. Fiscus, M. Michel et al. 2016. Trecvid 2016: Evaluating video search, video event detection, localization, and hyperlinking. In *Proceedings of TRECVID*, Vol. 2016.

[6] G. Ye, Y. Li, H. Xu, et al. 2015. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 471–480.

[7] M. Norouzi, T. Mikolov, S. Bengio et al. 2013. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650* (2013).

[8] O. Russakovsky, J. Deng, H. Su et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[9] P. Over, G. Awad, M. Michel et al. 2015. TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proceedings of TRECVID 2015*. NIST, USA.

[10] S. Wu, S. Bondugula, F. Luisier et al. 2014. Zero-shot Event Detection using Multi-modal Fusion of Weakly Supervised Concepts. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conf. on*. IEEE, 2665–2672.

[11] C. Xiaojun, Y. Yi, L. Guodong, Z. Chengqi, H. Alexander G. 2016. Dynamic concept composition for zero-example event detection. (2016), 20–26.

[12] Z. Fu, T. Xiang, E. Kodirov, S. Gong. 2015. Zero-shot object recognition by semantic manifold distance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2635–2644.

[13] L. Jiang, S.I. Yu, D. Meng, T. Mitamura, A.G. Hauptmann. 2015. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 27–34.

[14] L. Jiang, T. Mitamura, S.I. Yu, A.G. Hauptmann. 2014. Zero-example event search using multimodal pseudo relevance feedback. In *Proceedings of International Conference on Multimedia Retrieval*. ACM, 297.

[15] Ian Jolliffe. 2002. *Principal component analysis*. Wiley Online Library.

[16] Y.J. Lu. 2016. Zero-Example Multimedia Event Detection and Recounting with Unsupervised Evidence Localization. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1464–1468.

[17] E. Gabrilovich, S. Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis.. In *IJCAI*, Vol. 7. 1606–1611.

[18] C. Tzelepis, D. Galanopoulos, V. Mezaris, I. Patras. 2016. Learning to detect video events from zero or very few video examples. *Image and vision Computing* 53 (2016), 35–44.

[19] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, M. Shah. 2012. High-level event recognition in unconstrained videos. *International Journal of Multimedia Information Retrieval* (2012), 1–29.

[20] A. Habibian, T. Mensink, C. Snoek. 2014. Composite concept discovery for zero-shot video event detection. In *Proc. of Int. Conf. on Multimedia Retrieval*. ACM, 17.

[21] A. Habibian, T. Mensink, C. Snoek. 2014. VideoStory: A New Multimedia Embedding for Few-Example Recognition and Translation of Events. In *Proc. of the ACM Int. Conf. on Multimedia*. ACM, 17–26.

[22] P. Mettes, D. Koelma, C. Snoek. 2016. The ImageNet Shuffle: Reorganized Pre-training for Video Event Detection. *arXiv preprint arXiv:1602.07119* (2016).