T2R2 東京工業大学リサーチリポジトリ
Tokyo Tech Research Repository

論文 / 著書情報
Article / Book Information

論題(和文)	
Title(English)	Concept Elimination for Zero-Shot Event Detection
著者(和文)	ﾁｬﾝ ﾀﾞﾝ ﾊｲ, 井上 中順, 篠田 浩一
Authors(English)	Tran Hai Dang, Nakamasa Inoue, Koichi Shinoda
出典(和文)	第22回画像センシングシンポジウム (SSII) 講演論文集, , , IS2-19
Citation(English)	, , , IS2-19
発行日 / Pub. date	2016, 6
Note	第22回画像センシングシンポジウム (SSII), 2016講演論文集より転載

Powered by T2R2 (Tokyo Institute Research Repository)

Concept elimination for zero-shot event detection

Tran Hai Dang†, Nakamasa Inoue†, and Koichi Shinoda†
†Tokyo Institute of Technology
E-mail: {dang,inoue}@ks.cs.titech.ac.jp, shinoda@cs.titech.ac.jp

Abstract

We propose a concept selection algorithm for zero-shot event detection. A major approach to zero-shot event detection, which detects an event specified by a textual query from a video database, is to utilize a set of concept detectors built by using the other database. For example, to detect an event "Attempting a bike trick", detectors of concepts such as "bike", "helmet" are selected and utilized. Our proposed method not only finds the textually similar concepts, but also validate them based on senses of meaning. In the experiments on the TRECVID dataset, our method improved the APs in 6 of 20 events.

Figure 1 Example events in the TRECVID dataset.

1 Introduction

Recently, the fast development of video services on the Internet has led the demand for effective techniques for video search. As a result, multimedia event detection (MED) [1], which aims to find an user-defined *event*, is receiving research attention.

In MED, the detection using the relevant concepts is also useful for the detection of the whole event. For example, the event detection of "Attempting a bike trick" (Figure 1) can use concepts such as "bike", "helmet", etc. Therefore, the selection of the right concepts is an important task. We propose a method to deal with this problem.

We first choose concepts from a database of concept detectors using textual similarity, and then we eliminate irrelevant concepts based on the senses of meaning of words. Specifically, word2vec tool [2] with Google News dataset is used to produce a space of word vectors, and similarity between words is measured by cosine similarity on this space.

From a database of concept detectors that is built from datasets such as FCVID [3], Sports-1M [4], the concepts that are most similar to the event are selected. In order to avoid the misunderstanding concepts that will degrade the MED system, WordNet [5] is used to discover synonyms of words in the event, and the number of occurrences in Google n-gram is used to decide the right sense of meaning. After eliminating the misunderstanding concepts, the scores of the selected concept detectors are combined to evaluate the videos.

Our main contribution is the method that does not only find the textually similar concepts, but also validate them. These concept detectors help to increase the effectiveness of the MED task. We evaluate the system by conducting experiments on the TRECVID dataset [1]. Our method demonstrates advantages in events that contain multiple meaning words.

2 Related Work

For zero-shot event detection, a major approach is to use concept detectors from the other database, proposed by Ebadollahi *et al* [7] and Merler *et al* [8]. Here, a concept is an object, a scene, or an action. Some

previous studies have investigated effective concept databases. For example, Habbibian *et al* [9] pointed out that a database should consists more than 200 concepts of both specific and generic. Liu *et al* [10] proved that only a few of appropriate concepts for each event are important. This shows that selecting the relevant concepts for each event is an important problem.

To estimate the relevance, word similarity is often used. Typically, word vector representation such as word2vec [2] is introduced to measure word similarity [11] [12] [13]. The word2vec gives a vector representation of each word by using a neural network called skip-gram [14] trained on text corpus. Notably, it is used in the system in [13], which performed the best at the TRECVID 2015 Multimedia Event Detection task, to select the relevant concepts.

3 Proposed Method

The overview of our proposed method is shown in Figure 2. First, concepts are selected based on word similarity from a given event (query). Second, our concept elimination algorithm is applied to validate these concepts. Third, event detection scores are computed for each video by using concept detectors. Here, concept detectors are assumed to be trained on video data. The following subsections present details of each step.

3.1 Concept Selection

Let E be an event given as a set of words, and $\mathcal{C}$ be a set of concepts. We select pairs of a concept $c \in \mathcal{C}$ and a word $w \in E$ if the cosine similarity between corresponding word vectors is larger than a predefined threshold σ_1. Here, word vectors are extracted by using the word2vec tool [2].

3.2 Concept Elimination

In the concept elimination step, irrelevant concepts are eliminated based on senses of meaning as follows.

1. List all senses of meaning that have a set of synonyms of word w from WordNet. We skip senses of meaning that have no synonym.
2. Score each synonym s by the number of occurrences of the word sequence of the event name, which replaces w by s, in Google n-gram.
3. Select the sense of meaning that contains the synonym with the highest score.
4. Eliminate the pair (c, w) if the average of the similarities between the concept c and each synonym of the selected sense of meaning for the word w is lower than a predefined threshold σ_2.

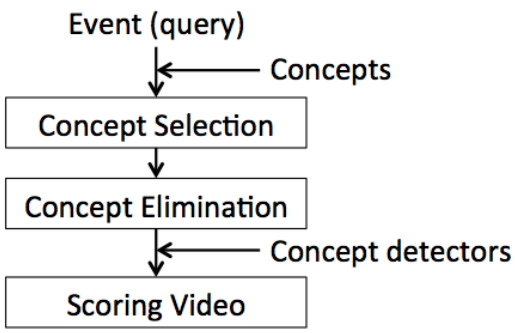

Figure 2 Overview of our method.

3.3 Scoring

Finally, we compute the event detection score from the scores of selected concept detectors by

$$S_E(X) = \frac{\sum_{c \in \Omega(E)} \alpha(E, c) f_c(X)}{\sum_{c \in \Omega(E)} \alpha(E, c)}, \qquad (1)$$

where $\Omega(E)$ is the set of relevant concepts of an event E, $f_c(X)$ is a score for a concept c, and $\alpha(E, c)$ is the average similarity between the concept c and each synonym corresponding to E given in the Step 4 in the algorithm presented in Section 3.1.

The score $f_c(X)$ is computed by a statistical model such as a support vector machine (SVM) for each concept. For example, SVMs trained with Gaussian mixture model (GMM) supervectors of several types of low-level visual or audio features can be introduced.

4 Experiments

4.1 Experimental setting

In our experiments, TRECVID Multimedia Event Detection (MED) dataset [1] is used. We use 12,632 videos (Kindered subset) and 20 types of events (MED14PS definition). Note that training data is not given for zero-shot event detection.

We build a concept database from the following datasets:

1. Fudan-Columbia Video Dataset (FCVID). This dataset consists of 91,223 videos collected from YouTube and Vimeo archives for 239 concepts. The total duration is 4,232 hours.
2. Sports-1M Dataset. This dataset consists of 1,133,158 videos collected from YouTube for 487 concepts. We manually select 20 concepts of actions and object.

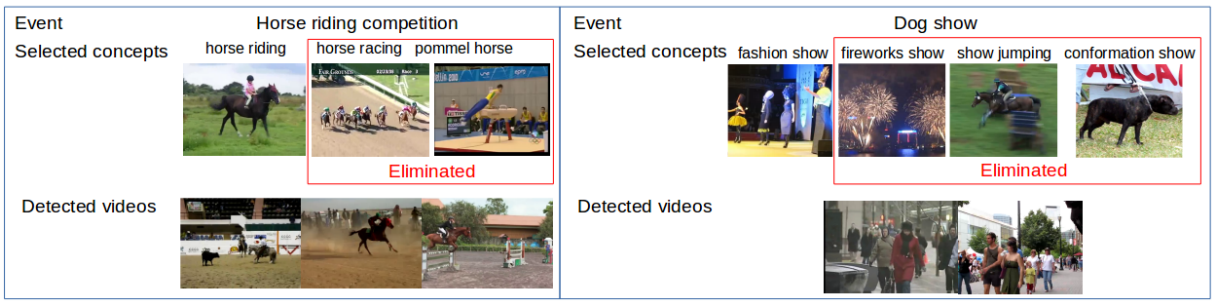

Figure 3 Example of eliminated concepts for events "Horse riding competition" and "Dog show"

Concept detectors are trained with GMM supervectors [15] of dense trajectories (DT) with MBH descriptors (DT-MBH) [16]. Word vectors are extracted by using the word2vec tool from the Google News dataset, which contains 100 billion words. We use pretrained word vectors for 3 million unique words. It has 117,000 synsets of English words. Google n-gram [6] is to count the occurrences of word sequences. Here we set the thresholds $\sigma_1 = 0.7, \sigma_2 = 0.5$. As a baseline, we utilize the concept selection method proposed in [13]. We also conduct a run that selects concept by only setting a threshold of similarities.

The evaluation measure is mean average precision (Mean AP), the geometric mean of AP among all 20 events. Average precision (AP) for each event is given by

$$AP = \frac{1}{R}\sum_{r=1}^{N} Pr(r)Rel(r), \tag{2}$$

where N is the number of testing videos, R is the number of positive videos, $Pr(r)$ is the precision at the rank r, and $Rel(r) \in \{0, 1\}$ is the positive or negative label of the r-th video.

4.2 Results

Our proposed method improved the mean AP from 7.39% to 10.59% with the performance improvement for 6 of 20 events as shown in Figure 4. For these 6 events, we confirmed that our algorithm successfully eliminated irrelevant concepts. For example, for an event "Horse riding competition","pommel horse" is eliminated as shown in Figure 3

The rest 14 events can be grouped into two: irrelevant concepts are not found and relevant concepts are eliminated. For the first group, all concepts selected based on word similarity were relevant, e.g., concepts "cleaning carpet" and "cleaning floor" for "Cleaning an appliance". For the second group, our algorithm failed to select the right sense of meaning. For example, for "Dog show", a concept "conformation show" was wrongly eliminated since the word conformation is not directly related to dogs. Dictionary of phrase is needed to improve the performance for such events.

5 Conclusion

In this study, we have proposed a method to eliminate irrelevant concepts based on senses of meaning after concept selection based on word similarity. Our method improved the performance of zero-shot event detection, for 6 of 20 events in the TRECVID dataset. Mean AP was improved by 3.2. Our future work will focus on dealing with not only separate words, but also phrases in the event.

References

[1] P. Over, J. Fiscus, G. Sanders, D. Joy, M. Michel. TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. *Proc. TRECVID*, 2014.

[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient Estimation of Word Representations in Vector Space. In *Proc. ICLR*, 2013.

[3] Y.-G. Jiang, Z. Wu, J. Wang, X. Xue, and S.-F. Chang. Exploiting Feature and Class Relationships in Video Categorization. In *arXiv preprint arXiv:1502.07209*, 2015

[4] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale Video Classification with Convolutional Neural Networks. In *Proc. CVPR*, 2014.

[5] G. A. Miller. WordNet: A Lexical Database for English. In *Communications of the ACM*, 1995.

[6] T. Brants, and A. Franz. Web 1T 5-gram Version 1 LDC2006T13. DVD. Philadelphia: Linguistic Data Consortium, 2006.

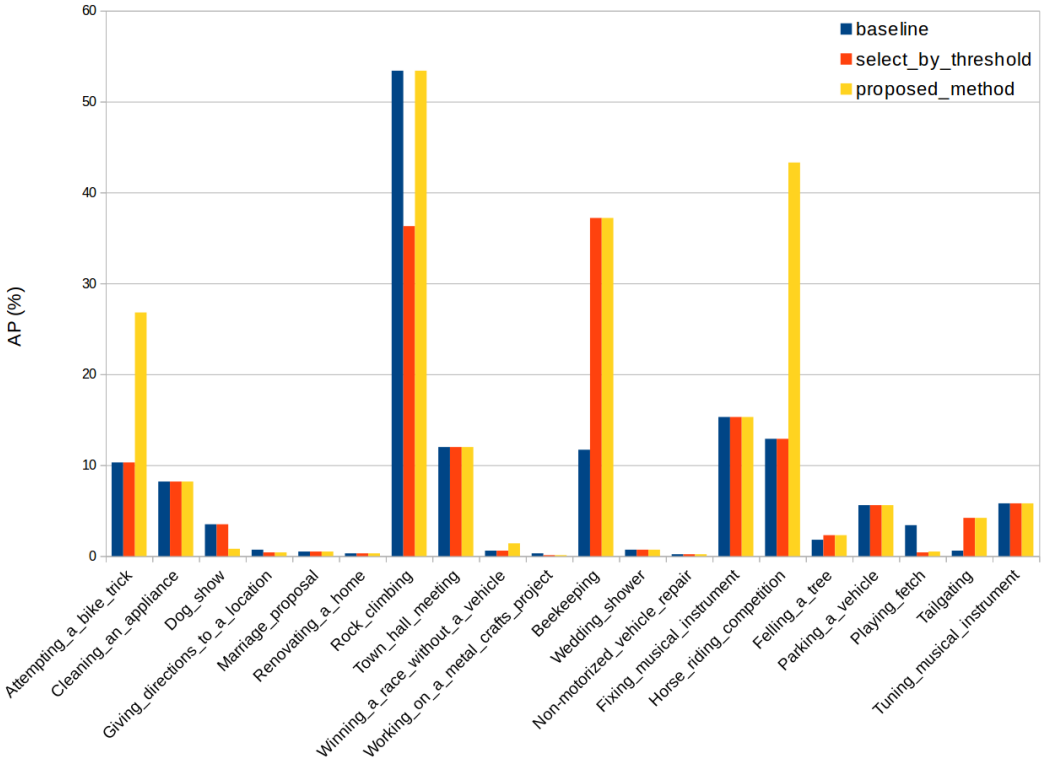

Figure 4 Experimental results of 20 events in AP (%)

[7] S. Ebadollahi, L. Xie, S.-F. Chang, and J. R. Smith. Visual event detection using multi-dimensional concept dynamics. In *Proc. ICME*, 2006.

[8] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. In *IEEE Trans. Multimedia*, 2012.

[9] A. Habibian, K. E. A. van de Sande, and C. G. M. Snoek. Recommendations for video event recognition using concept vocabularies. In *Proc. ICMR*, 2013.

[10] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. S. Sawhney. Video event recognition using concept attributes. In Proc. IEEE Workshop on Applications of Computer Vision, 2013.

[11] S. Yu, L. Jiang, Z. Mao, X. Chang, X. Du, C. Gan, Z. Lan, Z. Xu, X. Li, Y. Cai, A. Kumar, Y. Miao, L. Martin, N. Wolfe, S. Xu, H. Li, M. Lin, Z. Ma, Y. Yang, D. Meng, S. Shan, P. D. Sahin, S. Burger, F. Metze, R. Singh, B. Raj, T. Mitamura, R. Stern, and A. Hauptmann. Informedia @ TRECVID 2014. In *Proc. TRECVID Workshop*, 2014.

[12] C. Ngo, Y. Lu, H. Zhang, T. Yao, C. Tan, C. Tan, L. Pang, M. Boer, J. Schavemaker, K. Schutte, and W. Kraaij. VIREO-TNO @ TRECVID 2014: Multimedia Event Detection and Recounting (MED and MER). In *Proc. TRECVID Workshop*, 2014.

[13] H. Zhang, Y. Lu, M. de Boer, F. ter Haar, Z. Qiu, K. Schutte, W. Kraaij, and C. Ngo. VIREO-TNO @ TRECVID 2015: Multimedia Event Detection. In *TRECVID Workshop*, 2015.

[14] D. Guthrie, B. Allison, W. Liu, L. Guthrie, and Y. Wilks. A Closer Look at Skip-gram Modelling. In *Proc. LREC*, 2006.

[15] Y. Kamishima, N. Inoue, and K. Shinoda. Event detection in consumer videos using GMM supervectors and SVMs. *EURASIP Journal on Image and Video Processing*, 2013.

[16] H. Wang, and C. Schmid. Action Recognition with Improved Trajectories. In *Proc. ICCV*, 2013.

[17] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Proc. NIPS*, 2013.

[18] N. Inoue, T. H. Dang, R. Yamamoto, and K. Shinoda. TokyoTech at TRECVID 2015. In *TRECVID Workshop*, 2015.