Video Event Detection by Exploiting Word Dependencies from Image Captions

Sang Phan[†], **Yusuke Miyao**[†], **Duy-Dinh Le**[†§], **Shin'ichi Satoh**[†]
[†]National Institute of Informatics, Japan
[§]Multimedia Communications Lab, University of Information Technology, Vietnam
{plsang,yusuke,ledduy,satoh}@nii.ac.jp

Abstract

Video event detection is a challenging problem in information and multimedia retrieval. Different from single action detection, event detection requires a richer level of semantic information from video. In order to overcome this challenge, existing solutions often represent videos using high level features such as concepts. However, concept-based representation can be confusing because it does not encode the relationship between concepts. This issue can be addressed by exploiting the co-occurrences of the concepts, however, it often leads to a very huge number of possible combinations. In this paper, we propose a new approach to obtain the relationship between concepts by exploiting the syntactic dependencies between words in the image captions. The main advantage of this approach is that it significantly reduces the number of informative combinations between concepts. We conduct extensive experiments to analyze the effectiveness of using the new dependency representation for event detection on two large-scale TRECVID Multimedia Event Detection 2013 and 2014 datasets. Experimental results show that i) Dependency features are more discriminative than concept-based features. ii) Dependency features can be combined with our current event detection system to further improve the performance. For instance, the relative improvement can be as far as 8.6% on the MEDTEST14 10Ex setting.

1 Introduction

Detecting event from videos has been an important research topic due to the explosion of internet videos. In order to build a reliable event detection system, one must rely on the video content rather than simply using textual metadata (Davidson et al., 2010). However, internet videos are often captured under arbitrary conditions, which makes the large content variation among videos of the same event. To handle this problem, we can represent videos using multimodal features such as Dense trajectories (Wang and Schmid, 2013), SIFT (Lowe, 2004), and audio MFCC (Lee et al., 1988). Another approach to tackle this problem is to represent videos using the concept-based representation, i.e., video is represented by the concept detection scores indicating the presence of the concepts. This approach is particularly helpful when the number of training videos are scarce (Chen et al., 2014; Habibian et al., 2014b; Habibian et al., 2013; Ma et al., 2013; Ye et al., 2015). Some other works further select a subset of informative concepts for event detection (Jiang et al., 2015; Mazloom et al., 2013).

Nevertheless, the concept-based representation has two drawbacks i) It is non-trivial to obtain a large concept vocabulary. That is why people often combine concepts from multiple vocabularies to construct a larger one, in the hope to cover a wide range of concepts in real world videos (Yu et al., 2014). ii) The relationship between concepts, e.g., co-occurrence, are not captured. However, these relationships can convey a richer level of semantic information which can not be found from encoding individual concepts. Figure 1 illustrates the benefit of exploiting these relationships for event detection.

One can obtain the relationships between concepts by harvesting the social-tagged images (Li et al., 2012), however, tag information is often far from the actual content of the image. Visual phrases (Sadeghi and Farhadi, 2011) models a person and an interacting object through the use of a composite template.

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3318–3327, Osaka, Japan, December 11-17 2016.

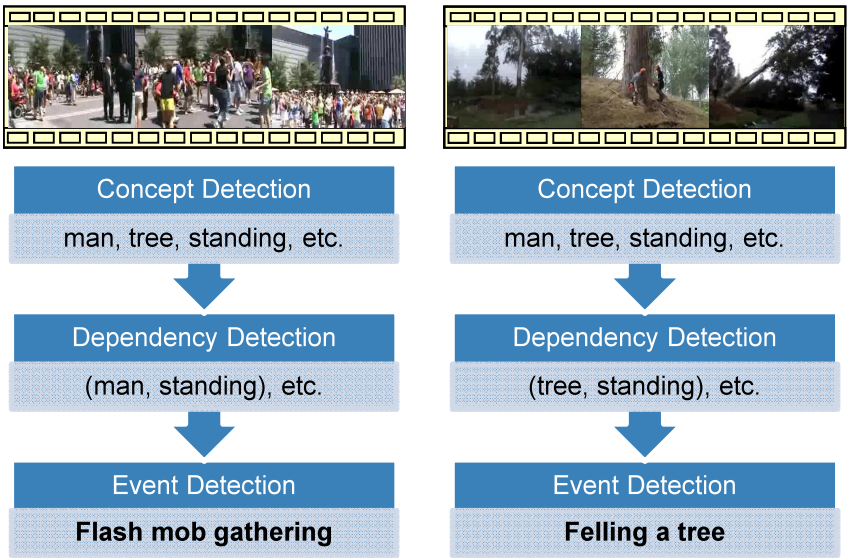

Figure 1: An illustrative example of limitation when using concept-based representation for video. In both videos, "man", "tree", and "standing" can be detected. However, if the dependency *(man, standing)* is also detected, it is more likely a "flash mob gathering". On the other hand, if *(tree, standing)* is detected, it is probably a "Felling a tree" event.

This approach might not be applicable to real world application because there are only 17 visual phrases being used and they are manually selected. Singh et al. (Singh et al., 2015) select relevant concept pairs for event detection by discovering from the event text query. Habibian et al. (Habibian et al., 2014a) also utilize the event query to discover the co-existence or exclusive existence relationships for zero-shot event detection. Assari et al. (Assari et al., 2014) and Can et al. (Can and Manmatha, 2014) model the concept co-occurrences or dependencies based on the joint distributions of two concepts, which relies on the concept detection scores. Borth et al. (Borth et al., 2013) have exploited some syntactic dependencies, such as the combination of adjectives and nouns, for training a sentiment analysis model.

The main drawback of the existing works is that the concept relationships are either manually defined or learned from the concept-based representation, which may face the effect of cumulative error. We address the former issue by exploiting more relationships between concepts such as syntactic dependency relations like subject and object. Such kind of syntactic relationships can also automatically eliminate irrelevant combinations between words or concepts. Typically, we propose a new approach to obtain the relationship between concepts by extracting the word dependencies from the image captions using standard natural language processing technologies. We then train a dependency model directly from the captioned images and use these dependencies as features for event detection.

What differentiates our approach from the previous works is that we explore a comprehensive set of relationship between concepts that is based on the syntactic dependencies. The immediate benefit of our approach is that we can easily obtain a large and informative dependency vocabulary from a much smaller concept vocabulary. In short, the contributions of this paper are twofold: i) A pilot study to exploit word dependencies as features for video event detection. These dependencies encode not only the co-occurrences but also various types of co-occurrences between concepts. ii) We report comprehensive experiments to demonstrate the benefit of the new dependency-based representation.

2 Word Dependency

The central idea in the present paper is to exploit word dependencies as features for event detection. Therefore, the development of the model to predict dependencies for a given video is a key issue. In the following section, we describe a method for obtaining training data for dependency prediction, and our model for dependency prediction.

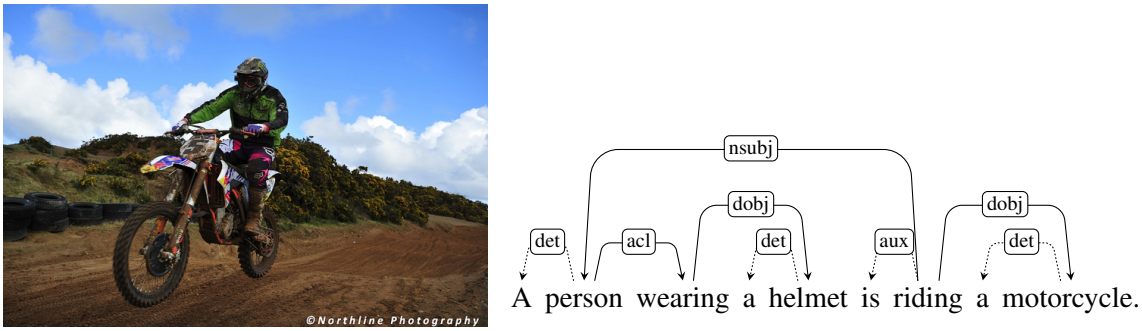

Figure 2: Example of word dependencies. Solid arrows show dependencies included in our vocabulary, while dotted arrows indicate dependencies excluded.

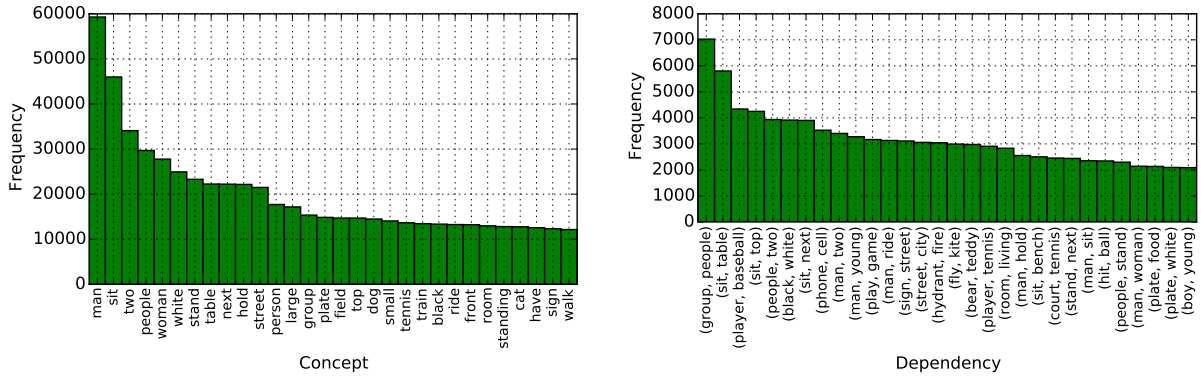

Figure 3: Top 30 frequent concepts and dependencies extracted from the MSCOCO captions.

2.1 Dependency Extraction

Dependency tree represents relationships among words in a sentence, such as subject-verb and verb-object relations. Figure 2 shows an image, its caption, and the dependency tree of the caption obtained by applying Stanford Parser (Manning et al., 2014). For example, an arrow from "riding" to "person" is labeled with "nsubj", which means "person" is a *nominal subject* of "riding".[1] Dependency trees can represent more precise semantic information than a bag of concepts. In this example, the dependecy tree indicates that "person" is the subject of "riding", not the object. This distinction is crucial in event detection as exemplified in Figure 1.

Dependency tree is defined as a set of dependencies, each of which is a triple (*label*, *head*, *dependent*), where *label* is a dependency label (e.g. "nsubj"), *head* is a source word of the dependency (e.g. "riding"), and *dependent* is the other side of the dependency (e.g. "person").

As dependencies consist of pair of words, the simplest way to construct a vocabulary of dependecies is to enumerate all pairs of words. However, this is obviously infeasible. When we have n words and k dependency labels, the total number of dependencies is kn^2. As in the typical situation where $n = 20,000$ and $k = 40$, we need to consider 16 billion dependencies.

In fact, most of the dependencies are not useful for event detection. One reason is that dependencies that appear in real text are very sparse; i.e., most of the triples do not appear. For example, the dependency ("dobj", "wear", "motorcycle"), which means *something is wearing a motorcycle*, is unlikely to appear in real text, because it does not make sense. The distribution of dependencies is highly skewed, and it is not clever to consider all combinations of words and dependency labels. Another reason is that several types of dependencies represent only grammatical relations, and do not convey semantic information.

[1]Dependency labels are defined in the guideline of Universal Dependencies. For details, see: `http://universaldependencies.org/`.

Table 1: Comparison of our model prediction with human evaluation on the COCO validation dataset.

Mean AP	Model prediction	Human evaluation
Concept	0.4347	**0.4410**
Dependency	0.1282	**0.2033**

For example, the dependency ("aux", "riding", "is") describes that "is" is an auxiliary verb of "riding", but this dependency is purely grammatical and does not include any semantic information.

These observations lead us to the idea that we extract a vocabulary of semantically informative dependencies from real texts by excluding unnecessary dependencies. Our solution here is to process dependencies obtained by dependency parsing in the following manner.

1. Apply a dependency parser to caption texts and obtain dependency trees.
2. Remove function words, such as determiners and punctuations, and select top n frequent words as a concept vocabulary.
3. Select dependencies in which both words are included in the concept vocabulary.
4. Cluster dependency labels that have similar relations.

Because our dependencies are extracted from real texts, our vocabulary of dependencies does not include semantically meaningless dependencies. Step 2 and 3 further exclude purely grammatical dependencies and less frequent dependencies. Step 4 is intended to further reduce semantically subtle distinctions. For example, two labels, "nsubj" and "csubj", are defined to represent subject relations, but their distinction (nominal vs. clausal) is not particularly important for our purpose. Therefore, we collapse such semantically similar dependency labels to get a reduced number of dependency types.

In Figure 2, solid arrows show dependencies included in our vocabulary, while dotted arrows indicate dependencies excluded. This example demonstrates that our method selects dependencies that capture semantic information of original texts while excluding less informative dependencies.

2.2 Dependency Modeling

Images and caption texts for the development of the dependency prediction model are obtained from the training set of Microsoft COCO (Lin et al., 2014), which contains 82,783 images and 414,113 captions constructed by crowdsourcing. Stanford CoreNLP 2015-04-20 is used for dependency parsing. We selected $n = 1,000$ concepts, which covers 89.4% of words (excluding function words) in the caption texts of the training data. Using this concept vocabulary, the method described in Section 2 extracted 20,931 dependencies that have more than 10 occurrences in the training data, which cover 63.3% of all the dependencies in the original caption texts. Top 30 frequent concepts and dependencies extracted from the training captions are shown in Fig. 3.

We consider the problem of predicting dependencies as a multi-label classification task because multiple dependencies can present in one image. To train the dependency model, we finetune a deep neural network on the COCO images from the VGG (16 layers) network (Simonyan and Zisserman, 2014). We add a cross-entropy loss layer on top of a sigmoid layer to account for the loss function. The entire network is trained using the Stochastic Gradient Descent (SGD) optimizer.

We compare the model predictions with the human performance to evaluate the effectiveness of our concept and dependency models. In order to measure the human accuracy, we use concepts/dependencies of one gold caption as the predicted labels, and use concepts/dependencies from the remaining captions of the same image as gold labels. If an image has multiple captions, we repeat this procedure so that every caption is used as the prediction, and report the average performance.

The detailed comparison of our model prediction with human performance is shown in Table 1. We found that our concept prediction model has comparable performance with agreement among human-annotated captions, while our dependency prediction is inferior to the human performance. The low agreement between annotators on the dependency evaluation also demonstrates the inherent challenge of dependency modeling.

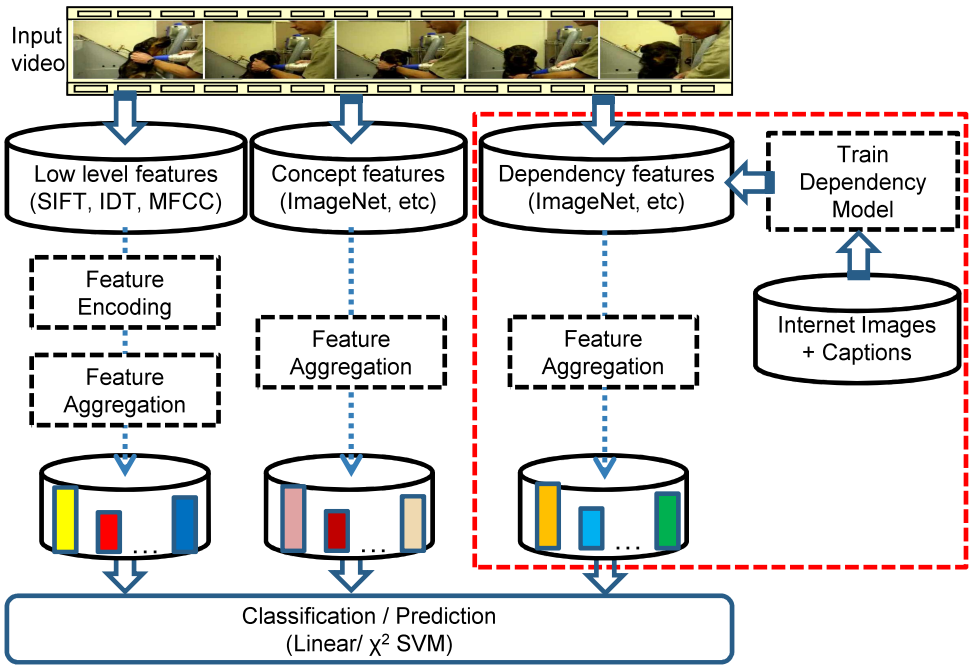

Figure 4: Overview of our event detection framework.

3 Event Detection Framework

In this section, we present a common event detection system, which consists of four main components: feature extraction, feature encoding, feature classification, and feature fusion (Fig. 4).

3.1 Feature extraction

Our framework supports a large variety of features including low-level and high-level features. Low-level features such as audio MFCC and Dense trajectories (Wang and Schmid, 2013) can be extracted from temporal windows spanning of several frames, while still image features such as SIFT (Lowe, 2004) is extracted from sampled frames of video. These raw low-level features require a special encoding technique to generate video-level representation, which will be described in Section 3.2.

We also extract high-level features such as visual concept features extracted from deep visual models. For example, our framework supports state-of-the-art deep learning features which are extracted from pre-trained deep models (Simonyan and Zisserman, 2014). These models are trained on large image collection such as ImageNet (Deng et al., 2009). Using these models, we can extract concept scores for each sampled frame and aggregate them into the final representation for each video. One can also utilize features from the previous fully-connected layers, such as `fc6` and `fc7`, to train the event detectors because they often retain richer information than the last full-connected layer.

Our concept and dependency features can be also considered as high-level features. However, different from the aforementioned high-level features that are extracted from pre-trained visual models, our concept and dependency features are extracted from our concept and dependency models respectively.

3.2 Feature encoding

We use Fisher Vector encoding to map local descriptors extracted from low-level features into the video-level representation. The Fisher vector (FV) has been successfully employed for various image and action classification problems (Perronnin and Dance, 2007; Wang and Schmid, 2013). It can be considered as an extension of Bag-of-Words (BoW) encoding where both first- and second-order statistics between the local descriptors and their assigned codewords are also encoded. Therefore, Fisher vector can achieve comparable performance to that of BoW while using a much smaller codebook.

In our experiment, we first randomly selected one million local descriptors for training a Gaussian mixture model (GMM), which is later used as the codebook for encoding. As suggested in (Perronnin et

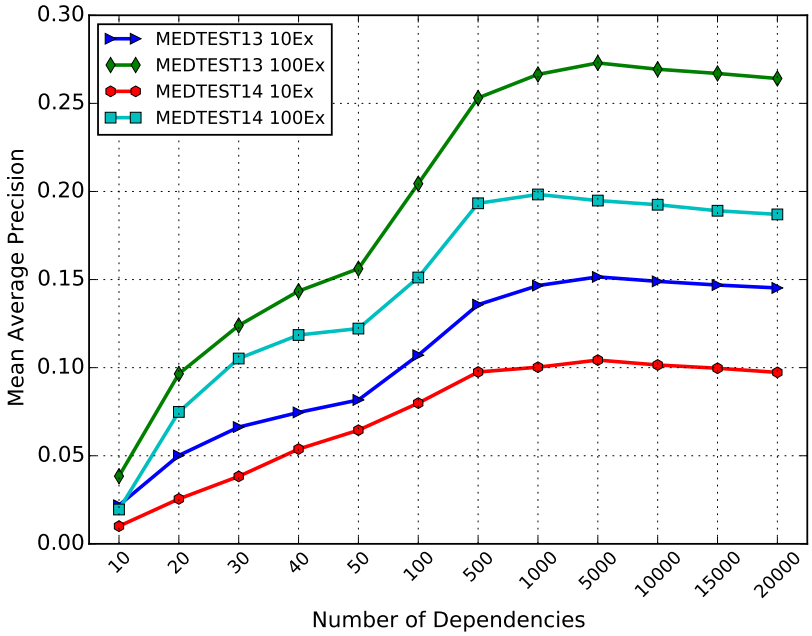

Figure 5: Results of event detection performance with respect to various vocabulary sizes.

al., 2010), it is better to reduce the local feature dimension by using principal component analysis (PCA) and apply power normalization, e.g., with $\alpha = 0.5$, followed by L2-normalization to the Fisher vector.

3.3 Feature classification

We use LibSVM (Chang and Lin, 2011) for training event detectors. The pre-computed kernel technique is utilized to reduce the training time. This technique is especially useful when the number of events are large. For low-level features, which are encoded using Fisher vector encoding, we use linear SVM for training and testing. For high-level features, since its feature dimension is rather small, we employ the χ^2 kernel SVM to obtain a better recognizing performance.

3.4 Feature fusion

Different features cover various characteristics of multimedia data. Hence it is natural to combine these features to get the benefit from multi-modal features. For the sake of simplicity, we use the average late fusion strategy for feature combination.

4 Experiments

We conduct experiments on the TRECVID Multimedia Event Detection (MED) 2013 and 2014 benchmarks under the 10Ex and 100Ex settings, in which only 10 and 100 training videos are given for each event respectively (Over et al., 2014). Dependency features are extracted from the final layer (fc8) of the dependency network presented in Section 2.2 for each sampled frame in video (1 frame every 4 seconds) and then aggregated to form the video representation. We employ LibSVM (Chang and Lin, 2011) with the exponential-χ^2 kernel for event training and testing. All the results are reported in terms of mean average precision (mean AP).

4.1 How many dependencies?

We first study how different number of dependencies affect to the event detection performance. We found that selecting the most frequent dependencies contributes more than randomly selecting the same number of dependencies. Following this observation, we only consider the top frequent dependencies, and report the results in Fig. 5. Small vocabulary sizes often result in a poor outcome. When increasing the vocabulary size, the performance also increases rapidly. The highest performance can be obtained

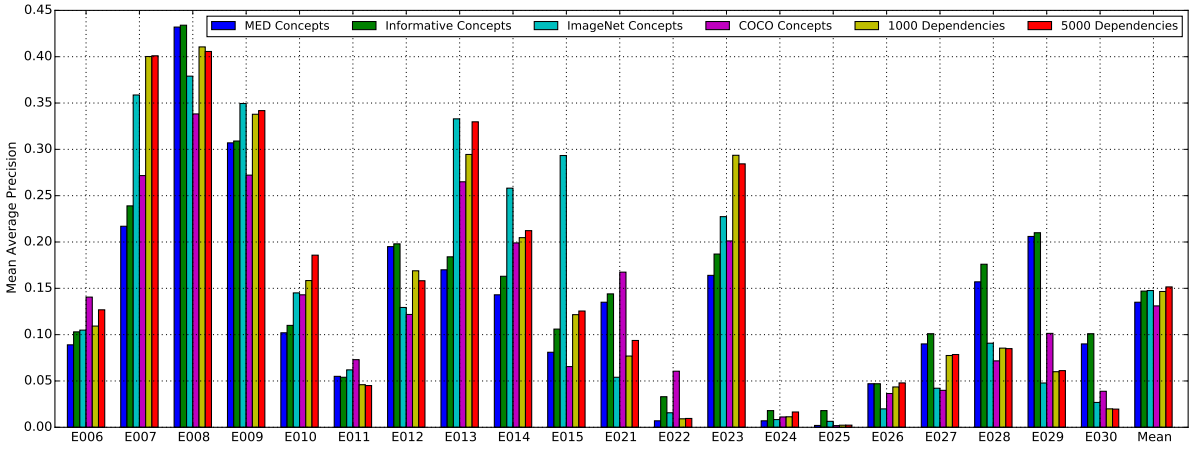

Figure 6: Performance comparison of dependency features with other methods on the MEDTEST13 10Ex setting. Performance of each event is reported in each group, the last group shows the average performance over all events. *MED Concepts* refers to concepts obtained from video annotations (Habibian et al., 2013), and *Informative Concepts* refers to selected concepts in (Mazloom et al., 2013). The mean average precisions from left to right are 0.1350 (MED Concepts), 0.1470 (Informative Concepts), 0.1476 (ImageNet Concepts), 0.1310 (COCO Concepts), 0.1466 (1000 Dependencies) and **0.1515** (5000 Dependencies).

by using around thousands of dependencies. However, if more than ten thousands of dependencies are added to the vocabulary, the performance tends to slightly drop. The reason is that at this point, the dependencies become much less frequent, so their detectors might become less reliable as well.

4.2 Comparison with other methods

Figure 6 shows the performance of our dependency features on the MEDTEST13 10Ex. We include the reported results on the same setting in (Habibian et al., 2013) and (Mazloom et al., 2013) for comparison. The performance of ImageNet concepts (Deng et al., 2009) are produced using the same network that was used to fine-tune our dependency model (Simonyan and Zisserman, 2014). We also select 1,000 COCO concepts as described in Section 2 and report its performance. The best dependency run can outperform the COCO concept run with a relative improvement of **15.6%**. Our dependency features only achieve a marginal improvement over the ImageNet baseline. However, this result is already encouraging because our dependency model was trained with around 80K labeled images, compared to about one million labeled images of the ImageNet model.

4.3 Contribution to our MED system

Finally, we conduct experiments to demonstrate the contributions of our dependency feature to our event detection system. State-of-the-art event detection systems (Yu et al., 2014; Yu et al., 2015) often employ features from multiple modalities such as audio and visual features. Following this strategy, we implemented a number of features in our system including audio MFCC (Lee et al., 1988), SIFT (Lowe, 2004), and Dense trajectories (Wang and Schmid, 2013). In light of the success of deep learning, we also use features extracted from (Simonyan and Zisserman, 2014), which is a pre-trained network on ImageNet (Deng et al., 2009). We also include the reported performance of recent state-of-the art systems in (Yu et al., 2014) and (Yu et al., 2015) for comparison.

Table 2 compares performance of all aforementioned features. Our (5,000-dimensional) dependency feature outperforms ImageNet concepts in all settings, while still inferior to the performance of the Dense trajectories features (Wang and Schmid, 2013). When combining our existing system with the new dependency feature, we achieve a relative improvement of **6.8%** and **8.6%** on the MEDTEST13 10Ex and MEDTEST14 10Ex settings respectively. We report lower results than the ones reported in (Yu et al., 2014; Yu et al., 2015). The reason can be due to the fact that they included more concept features

Table 2: Contribution of the new dependency features to our event detection system.

	MEDTEST13		MEDTEST14	
	10Ex	100Ex	10Ex	100Ex
MFCC (Lee et al., 1988)	0.0440	0.1010	0.0449	0.0776
SIFT (Lowe, 2004)	0.0893	0.2235	0.0730	0.1796
IDT (MBH) (Wang and Schmid, 2013)	0.1550	0.2812	0.0937	0.2138
IDT (HOGHOF) (Wang and Schmid, 2013)	0.1743	0.3198	0.1184	0.2595
ImageNet (Simonyan and Zisserman, 2014)	0.1476	0.2632	0.1037	0.1930
Dependencies	0.1515	0.2729	0.1043	0.1948
Late fusion (w/o dependencies)	0.2420	0.4101	0.1707	0.3449
Late fusion (w/ dependencies)	**0.2584**	**0.4244**	**0.1853**	**0.3571**
DMSY (Yu et al., 2015)	0.2800	0.3860	0.2330	0.3260
CMU (Yu et al., 2014)	0.3130	0.4640	0.2850	0.4190

in their system such as Google Sports (Karpathy et al., 2014) and YFCC concepts (Thomee et al., 2015). Example of some detected concepts for each event can be found in Table 3.

5 Conclusions and Future Work

We exploited word dependencies as a new semantic video representation for recognizing complex events. Different from the existing works, this representation encodes the relationship between concepts based on the syntactic dependencies between words. Therefore it captures semantically richer information, which is crucial for video event detection. We demonstrated that the dependency-based representation is more discriminative than the concept-based representation. Moreover, it also helps improve the detection performance of our existing event detection system.

One limitation of the current work is that we exploited image captions rather than video captions. The main reason is that we have not found any large-scale video captioning corpus for this project. Also the accuracy of dependency prediction from video would be less accurate since video is more challenging to model. In the future work, we plan to extend this work to further modeling word dependencies from video captions.

Acknowledgements

This research was partially supported by JST, PRESTO.

References

Shayan Assari, Amir Zamir, and Mubarak Shah. 2014. Video classification using semantic concept co-occurrences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2529–2536.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*.

Ethem F Can and R Manmatha. 2014. Modeling concept dependencies for event detection. In *Proceedings of International Conference on Multimedia Retrieval*, page 289. ACM.

Chih-Chung Chang and Chih-Jen Lin. 2011. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Table 3: Top 3 dependencies detected by our system for each event in the MEDTEST13 collection.

ID	Event name	Top dependencies
E006	Birthday party	(group, people), (group, woman), (stand, front)
E007	Changing a vehicle tire	(black, white), (man, young), (man, ride)
E008	Flash mob gathering	(crowd, people), (group, people), (walk, down)
E009	Getting a vehicle unstuck	(black, white), (photo, black), (man, young)
E010	Grooming an animal	(black, white), (man, young), (sit, chair)
E011	Making a sandwich	(stand, front), (man, young), (black, white)
E012	Parade	(crowd, people), (group, people), (walk, down)
E013	Parkour	(black, white), (man, young), (building, tall)
E014	Repairing an appliance	(take, picture), (man, young), (black, white)
E015	Working on a sewing project	(man, young), (take, picture), (phone, cell)
E021	Attempting a bike trick	(black, white), (man, ride), (board, skate)
E022	Cleaning an appliance	(black, white), (man, young), (take, picture)
E023	Dog show	(crowd, people), (group, people), (black, white)
E024	Giving directions to a location	(black, white), (man, young), (building, tall)
E025	Marriage proposal	(man, young), (crowd, people), (group, people)
E026	Renovating a home	(man, young), (black, white), (stand, front)
E027	Rock climbing	(man, young), (black, white), (hold, up)
E028	Town hall meeting	(crowd, people), (group, people), (group, woman)
E029	Winning a race without a vehicle	(group, people), (crowd, people), (man, young)
E030	Working on a metal crafts project	(phone, cell), (man, young), (man, wear)

Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, and Shih-Fu Chang. 2014. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *Proceedings of International Conference on Multimedia Retrieval*.

James Davidson, Benjamin Liebald, Junning Liu, Palash Nandy, et al. 2010. The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE.

Amirhossein Habibian, Koen EA van de Sande, and Cees GM Snoek. 2013. Recommendations for video event recognition using concept vocabularies. In *Proceedings of the 3rd ACM conference on International conference on multimedia retrieval*, pages 89–96. ACM.

Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. 2014a. Composite concept discovery for zero-shot video event detection. In *Proceedings of International Conference on Multimedia Retrieval*, page 17. ACM.

Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. 2014b. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *Proceedings of the ACM International Conference on Multimedia*, pages 17–26. ACM.

Lu Jiang, Shoou-I Yu, Deyu Meng, Yi Yang, Teruko Mitamura, and Alexander G Hauptmann. 2015. Fast and accurate content-based semantic search in 100m internet videos. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 49–58. ACM.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Chin-Hui Lee, F.K. Soong, and Bing-Hwang Juang. 1988. A segment model based approach to speech recognition. In *International Conference on Acoustics, Speech, and Signal Processing, 1988. ICASSP-88.*, pages 501 –541 vol.1.

Xirong Li, Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. 2012. Harvesting social images for bi-concept search. *IEEE Transactions on Multimedia*, 14(4):1091–1104.

T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *European conference on computer vision*.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Zhigang Ma, Yi Yang, Zhongwen Xu, Shuicheng Yan, Nicu Sebe, and Alexander G Hauptmann. 2013. Complex event detection via multi-source video attributes. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL*, pages 55–60.

Masoud Mazloom, Efstratios Gavves, Koen van de Sande, and Cees Snoek. 2013. Searching informative concept banks for video event detection. In *Proceedings of International Conference on Multimedia Retrieval*.

Paul Over, Georges Awad, Martial Michel, Johnatan Fiscus, Greg Sanders, Wessel Kraaij, Alan F Smeaton, and Georges Quénot. 2014. Trecvid 2014- an overview of the goals. *TRECVID*.

Florent Perronnin and Christopher Dance. 2007. Fisher kernels on visual vocabularies for image categorization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the fisher kernel for large-scale image classification. In *European conference on computer vision*, pages 143–156. Springer.

Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1745–1752. IEEE.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Bharat Singh, Xintong Han, Zhe Wu, Vlad I Morariu, and Larry S Davis. 2015. Selecting relevant web trained concepts for automated event retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4561–4569.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2015. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*.

Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3551–3558.

Guangnan Ye, Yitong Li, Hongliang Xu, Dong Liu, and Shih-Fu Chang. 2015. Eventnet: A large scale structured concept library for complex event detection in video. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference*, pages 471–480. ACM.

S Yu, L Jiang, Z Mao, XJ Chang, XZ Du, C Gan, ZZ Lan, ZW Xu, XC Li, Y Cai, et al. 2014. Cmu-informedia@ trecvid 2014 multimedia event detection (med). In *TRECVID Workshop*.

Shoou-I Yu, Lu Jiang, Zhongwen Xu, Yi Yang, and Alexander G Hauptmann. 2015. Content-based video search over 1 million videos with 1 core in 1 second. In *Proceedings of International Conference on Multimedia Retrieval*.