Large Video Event Ontology Browsing, Search and Tagging (EventNet Demo)

Hongliang Xu, Guangnan Ye, Yitong Li, Dong Liu, Shih-Fu Chang
Department of Electrical Engineering, Columbia University, New York, NY 10027, USA
{hx2168, gy2179, yl3029, dl2713, sc250}@columbia.edu

ABSTRACT

EventNet is the largest video event ontology existent today, consisting of 500 events and $4,490$ event-specific concepts systematically discovered from the crowdsource forum like WikiHow. Such sources offer rich information about events happening in everyday lives. Additionally, it includes automatic detection models for the constituent events and concepts using deep learning with around $95K$ training videos from YouTube. In this demo, we present several novel functions of *EventNet*: 1) interactive ontology browsing, 2) semantic event search, and 3) tagging of user-loaded videos via open web interfaces. The system is the first in allowing users to explore rich hierarchical structures among video events, relations between concepts and events, and automatic detection of events and concepts embedded in user-uploaded videos in a live fashion.

Categories and Subject Descriptors

I.2.10 [**ARTIFICIAL INTELLIGENCE**]: Vision and Scene Understanding—*Video Analysis*

Keywords

EventNet Ontology Browsing, Event Search, Automatic Video Tagging, Structured Video Event Ontology

1. INTRODUCTION

Recent advances in computer vision and multimedia have shown impressive progresses in automatic recognition of objects, actions, and events [3, 4, 5]. Significant performance gains have been periodically reported thanks to the availability of large data sources and sophisticated learning models. However, a few major questions remain open: what classes should be included in the recognition task and what structural relations might exist among classes?

Motivated by these challenges, we have developed a systematic process to discover common events that are relevant to human activities (e.g., “wedding", “dance party") or procedural tasks (e.g., “cooking", “car repair"). We focus on events that can be manifested visually in the video data and are with a reasonable level of details. Using data mining and machine learning processes, our efforts have resulted in an EventNet event ontology [6], which consists of 500 high level events and $4,490$ semantic concepts relevant to the discovered events.

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2807973.

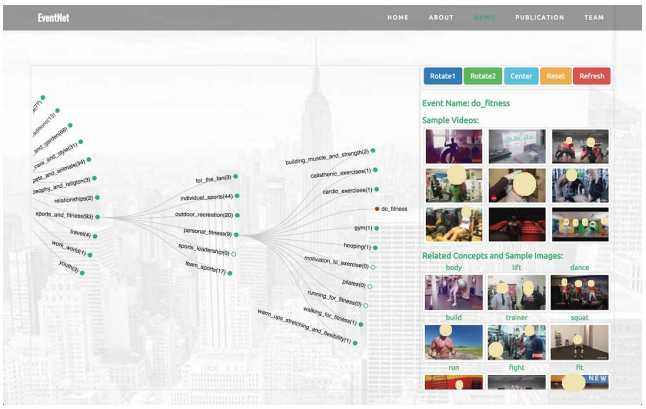

Figure 1: Interface for event ontology browsing. Example videos and related concepts of the selected event are shown.

In this demo, we present new functions of the EventNet system: interactive browser, semantic search, and live tagging of user-uploaded videos. In each of the modules, we emphasize the unique ontological structure embedded in EventNet and utilize it to achieve novel experience. For example, the event browser leverages the hierarchical event structure discovered from the crowdsource forum WikiHow [2] to facilitate intuitive exploration of events, the search engine focuses on retrieval of hierarchical paths containing interested events rather than events as independent entities, and finally the live detection module applies the event models and the associated concept models to explain why a specific event is detected in an uploaded video. To the best of our knowledge, the EventNet system is the largest video event ontology and the first interactive system allowing users to explore high level events and associated concepts in videos in a systematic structured manner.

2. BRIEF SUMMARY OF EVENTNET

EventNet ontology is discovered from WikiHow, an online forum teaching how to perform various how-to tasks in human daily life. Topics covered in such sources are also frequently captured in videos. WikiHow organizes topics into categories (currently $2,803$) using a hierarchical structure. The top layer contains 19 broad categories, and each category is further divided into subcategories that correspond to specialized subclasses or different facets of the parent category. The events in EventNet are found by first systematically inspecting articles in each category and then choosing an event that can be manifested by visual content (such as “bike riding", “apply makeup") . This process results in a total of 500

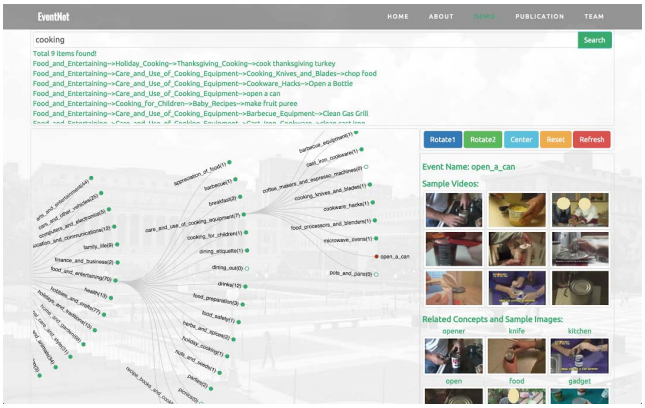

Figure 2: Interface for searching events embedded in the EventNet ontology.

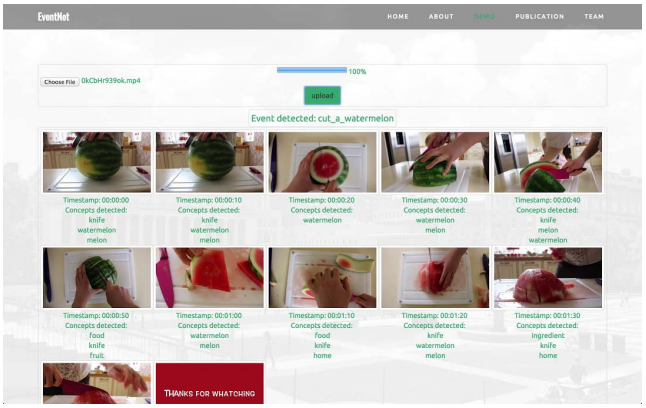

Figure 3: Interface of automatic tagging of user uploaded videos.

events, which are attached to the hierarchical ontology of WikiHow. For each event, we further use it as a textual query to search YouTube videos and discover event-specific concepts from the textual tags contained in the returned videos. This results in $4,490$ concepts in total. The pipeline mentioned above is systematic and productive, combining the rich information captured by the large crowdsource repository (WikiHow) about frequent events in human lives, and the almost unlimited video content available in online sources like YouTube.

With the large number of events and concepts, we crawl around $95K$ YouTube as training data for model learning. For event model, we use the video frames to train an AlexNet structure CNN model [4] over the 500 events. For concept model, we first extract the $4,096$-dimensional feature in the second last layer of the event model computed over each frame, and then train a binary SVM as the concept model. Details about EventNet construction can be seen in [6].

3. DEMO SYSTEM

Function I: Event Ontology Browser. Our system supports users to browse EventNet tree ontology in an interactive and intuitive manner. When a user clicks on a non-leaf category node, it expands the child category nodes together with any event attached to this category (the event node is filled in red while the category node is in green). When the user clicks on an event, it shows the exemplary videos of this event with a dynamic GIF showing animation of keyframes extracted from a sample video. Concepts specific to the event are also shown with representative keyframes of the concept. We specifically adopt the expandable, rotatable tree as the visualization tool (as shown in Figure 1) since it maintains a nice balance between the depth and breadth of the scope when the user navigates through layers and siblings in the tree.

Function II: Semantic Search of Events in the Ontology. We adopt a unique search interface that is different from conventional ones by allowing users to find hierarchical paths matching user interest, instead of treating events as independent units. This design is important for fully leveraging the ontology structure information in EventNet. For each event in EventNet, we generate its text representation by combining all words of the category names from root node to the current category containing the event plus the name of the event. Such texts are used to set up search indexes in Java Lucene [1]. When the user searches for keywords, the system will return all the paths in the index that contain the query keywords. If the query contains more than one words, the path that has more matched keywords will be ranked higher in the search result. After the search, users can click on each returned event, and our system will dynamically expand the corresponding path of this event and visualize it using the tree browser described in the previous section. This not only helps users quickly find target events, but also help suggest additional events to the user by showing events that may exist in the sibling categories in the EventNet hierarchy. Figure 2 shows the interface of the search function.

Function III: Automatic Video Tagging. EventNet includes an upload function that allows users to upload any video and use pretrained detection models to predict events and concepts present in the video. For each uploaded video, EventNet extracts one frame every 10 seconds. Each frame is then resized to 256 by 256 pixels and fed to the deep learning model described earlier. We average the 500-dimensional detection scores across all extracted frames and use the average score vector as the event detection scores of the video. To present the final detection result, we only show the top event with highest score as the event prediction of the video. For concept detection, we use the feature in the second last layer of the deep learning model computed over each frame, and then apply the binary SVM classifiers to compute the concept scores on each frame. We show the top-ranked predicted concepts under each sampled frame of the uploaded video. Figure 3 shows the tagging results of event and concept for a uploaded video, indicating the high accuracy of the tagging result. It is worth mentioning that our tagging system is very fast and satisfies real-time requirement. For example, when we upload 10 MB video, the tagging system can generate the tagging results in 5 seconds on a single regular workstation, demonstrating the high efficiency of the system.

4. CONCLUSION

This paper demonstrates novel functions on EventNet, the largest event ontology existent today with a hierarchical structure extracted from the popular crowdsource forum WikiHow. The system provides efficient event browsing and search interfaces, and supports live video tagging with high accuracy. It also provides a flexible framework for future scaling up by allowing users to add new event nodes to the ontology structure.

5. REFERENCES

[1] `https://lucene.apache.org/core/`.

[2] `http://www.wikihow.com/Main-Page`.

[3] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, M. Shah. High-level event recognition in unconstrained videos. In *IJMIR*, 2013.

[4] A. Krizhevsky, I. Sutskever, and G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*, 2012.

[5] S. Khurram, A. R. Zamir, and M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. In *arXiv:1212.0402*, 2012.

[6] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang. EventNet: A Large Scale Structured Concept Library for Complex Event Detection in Video. In *ACMMM*, 2015.