Event-Driven Semantic Concept Discovery by Exploiting Weakly Tagged Internet Images

Jiawei Chen*, Yin Cui*, Guangnan Ye, Dong Liu, Shih-Fu Chang
Department of Electrical Engineering, Columbia University
{jc3960,yc2776,gy2179}@columbia.edu, {dongliu,sfchang}@ee.columbia.edu

ABSTRACT

Analysis and detection of complex events in videos require a semantic representation of the video content. Existing video semantic representation methods typically require users to pre-define an exhaustive concept lexicon and manually annotate the presence of the concepts in each video, which is infeasible for real-world video event detection problems. In this paper, we propose an automatic semantic concept discovery scheme by exploiting Internet images and their associated tags. Given a target event and its textual descriptions, we crawl a collection of images and their associated tags by performing text based image search using the noun and verb pairs extracted from the event textual descriptions. The system first identifies the candidate concepts for an event by measuring whether a tag is a meaningful word and visually detectable. Then a concept visual model is built for each candidate concept using a SVM classifier with probabilistic output. Finally, the concept models are applied to generate concept based video representations. We use the TRECVID Multimedia Event Detection (MED) 2013 as our video test set and crawl $400K$ Flickr images to automatically discover $2,000$ visual concepts. We show significant performance gains of the proposed concept discovery method over different video event detection tasks including supervised event modeling over concept space and semantic based zero-shot retrieval without training examples. Importantly, we show the proposed method of automatic concept discovery outperforms other well-known concept library construction approaches such as Classemes and ImageNet by a large margin (228%) in zero-shot event retrieval. Finally, subjective evaluation by humans also confirms clear superiority of the proposed method in discovering concepts for event representation.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance

Keywords

Video Event Detection, Concept Discovery, Zero-Shot Retrieval, Semantic Recounting.

* indicates equal contributions.

ICMR '14 April 01 - 04 2014, Glasgow, United Kingdom

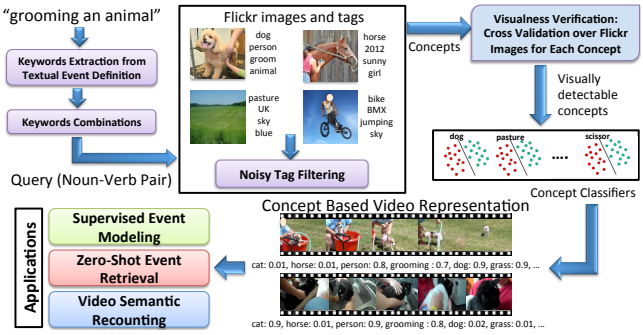


Figure 1: The framework of the proposed concept discovery approach. Given a target event (e.g., "grooming an animal") and its textual definition, we extract noun and verb keywords and use each noun-verb combination as a query to crawl images and their associated tags from Flickr. We then discover potential concepts from tags by considering their semantic meanings and visual detectability. We then train a concept model for each concept based on the Flickr images annotated with the concept. Applying the concept models to the videos generates concept-based video representations, which can be used in supervised event modeling over concept space, zero-shot event retrieval as well as semantic recounting of video content.

1. INTRODUCTION

Recognizing complex events from unconstrained videos has received increasing interests in multimedia information retrieval and computer vision research communities [12, 28, 30]. By definition, an event is a complex activity that involves people interacting with other people and/or objects under certain scene. Compared with human action recognition which focuses on simple primitives such as "jumping", "walking" and "running" [14, 17, 35], event detection is more challenging since it has to deal with unconstrained videos that contain various people and/or objects, complicated scenes and their mutual interactions. For example, a video of "birthday party" may contain a large number of atomic components including objects like "cake" and "candle", actions like "dancing" and "hugging" as well as scenes like "garden" and "living room".

The existing event detection works proposed the use of raw audio-visual features fed into different sophisticated statistical learning frameworks, and achieved satisfactory performance [13, 25]. However, these works are incapable of providing any interpretation or understanding of the abundant semantics presented in a complex multimedia event. This hampers high-level event analysis and understanding, especially when the number of training videos is small or zero. Therefore, a logical and computationally tractable way

is to represent a video depicting a complex event in a semantic space consisting of semantic concepts related to objects, scenes and actions, in which each dimension measures the confidence score of the presence of a concept in the video. Once we have such concept-based video representations, we will be able to use them as middle-level features in supervised event modeling or directly use the scores of the semantic concepts to perform zero-shot event retrieval [18, 23].

One intuitive approach to generate concept based video representation is to manually define a suitable concept lexicon for each event, followed by annotating the presence/absence of each concept in the videos [18, 22, 23]. Apparently, this approach involves tremendous manual efforts, and is impractical for real-world event detection problems regarding a huge number of videos. On the other hand, the web is a rich source of information with huge number of images captured for various events under different conditions, and these images are often annotated with descriptive tags that indicate the semantics of the visual content. Our intuition is that the tags of Internet images related to a target event should reveal certain common semantics appeared in the event, and thus suggest the relevant concepts in the event videos. This stimulates a challenging research problem which has not been well studied yet: given a target event (sometimes associated with a textual definition of the event, such as the textual event kits in TRECVID Multimedia Event Detection (MED) task [1]), how to automatically discover relevant concepts from the tags of Internet images, and construct corresponding concept detection models specifically optimized for the target events. Since we focus on discovering concepts for prespecified events, we term our approach as **Event-Driven Concept Discovery** to distinguish it from the work that builds a generic concept library independent of the targets [32]. Figure 1 illustrates the overall framework of the proposed system.

There are three main challenges in utilizing Internet images and their associated tags to learn concept models for complex event detection. First, since the tags associated with the Internet images are provided by general Internet users, there are often tags that are meaningless or irrelevant to the target event. To ensure the correctness of concept discovery, our method must choose semantic meaningful tags as the candidate concepts of the target event. To address this task, we perform noisy tag filtering by matching each tag to synsets in WordNet [24].

Moreover, some tags are abstract and not related to visual contents. For example, images with tags "economy" and "science" do not show consistent visual patterns that can be effectively modeled by computer vision techniques. Therefore, we employ a visualness verification procedure to check whether a tag can be visually detected. Only visually related tags are kept as candidate concepts.

The last challenge is that the labels of Internet images are often very noisy. Directly adopting such images as training samples for a concept will lead to a poor concept model. To solve this problem, we turn to a confidence ranking strategy. Given an image annotated with a concept, we first estimate the posterior probability of the concept's presence based on its visual closeness to other images annotated with the same concept. Then we rank all images based on the probabilities and choose the top ranked ones as the positive training samples of the concept. Such confidence ranking strategy is valuable in reducing the influence of noisy labels since it measures the confidence of the concept's presence in an image from its collective coherence with other images.

We will demonstrate both qualitatively and quantitatively that the proposed concept discovery approach is able to generate accurate concept representations on event videos. Applying the concept scores as concept representations of videos, our method achieve significant performance gains when evaluated over various semantic based video understanding tasks including supervised event modeling and zero-shot event retrieval. One major contribution is that the concepts discovered based on the proposed method achieves significant performance gains (228% in zero-shot event retrieval) over concept pools constructed using other well known methods such as Classemes and ImageNet. We also show that our discovered concepts outperform classic low-level features in supervised event modeling, and are able to reveal the semantics in a video over the semantic recounting task.

2. RELATED WORK

Complex event detection in videos has been investigated in literature. Duan *et al*. [7] proposed to learn cross-domain video event classifiers from the mixture of target event videos and source web videos crawled from Youtube. Tang *et al*. [31] developed a large margin framework to exploit the latent temporal structure in event videos, and achieved good performance on event detection. Natarajan *et al*. [25] exploited multimodal feature fusion by combining low-level features and available spoken and videotext content associated with event videos. Ma *et al*. [20] proposed to adapt knowledge from other video resources to overcome the insufficiency of the training samples in small sample video event detection. However, these works focus on modeling events into sophisticated statistical models, and cannot reveal the rich semantics in videos.

Some works attempted to accomplish event detection with concept based video representations. Izadinia *et al*. [11] manually annotated a number of concepts on event videos, and proposed a discriminative model that treats the concepts as hidden variables and models the joint relationship among concepts in a concept co-occurrence graph. Liu *at al*. [18] observed concepts in event videos and defined a concept ontology falling into "object", "scene", and "action", through which a number of SVM classifiers are trained as concept detectors to generate concept scores on videos. However, as aforementioned, all these methods require significant manual efforts, which are inadequate for real-world event detection task with a large number of videos. In [21], Ma *et al*. leveraged concepts contained in other video resources to assist detection in event videos, and proposed a joint learning model to learn concept classifier and event detector simultaneously. Mazloom *et al*. [22] first constructed a concept library with $1,346$ concept detectors by mixing 346 manually defined concepts in TRECVID 2011 Semantic Indexing Task [2] and $1,000$ concepts from the ImageNet Large Scale Visual Recognition Challenge 2011 [6], and then discovered the optimal concept subset using a cross-entropy optimization. Nevertheless, the concepts in other video resources may not be relevant to the content of event videos, and produce inaccurate semantic descriptions on videos. Yang *et al*. [38] adopted deep belief nets to learn cluster centers of video clips and treated them as data-driven concepts. Apparently, such data-driven concepts do not convey any semantic information, and are not applicable for semantic representation of videos. Differently, we focus on automatically discovering semantic concepts in event videos by exploiting the Internet images and their tags, which uncovers the semantics in videos without any manual labors.

Berg *et al*. [3] introduced a method to automatically discover concepts by mining text and image data sampled from the Internet. A text string is recognized as a concept only if the visual recognition accuracy on its associated images is relatively high. Nevertheless, the method merely works on a closed web image set with surrounding texts, and cannot be applied in concept discovery in event videos, each of which does not contain any textual description. Yanai *et al*. [37] adopted a similar idea to discover visual

(a) Attempting a bike trick

(b) Birthday party

Figure 2: Concept clouds for two example events, where the size of each concept indicates its TF-IDF value.

related concepts associated with Internet images. Our work is also related to analyzing videos by leveraging still images. For example, Ikizler-Cinbis *et al*. [10] proposed to learn actions from the web, which collected images from the Web to learn representations of actions and used this knowledge to automatically annotate actions in videos. In contrast, we focus on automatically discovering concepts from the still images and using them to interpret complex video semantics, which is more challenging than these prior works.

In terms of building a concept bank library, our work is also related to the existing concept libraries such as Object Bank [16], Classemes [32], and Action Bank [29]. However, these libraries are designed for generic objects or actions, and hence are not directly relevant to a target event collection at hand. Differently, our concept library is designed for a set of pre-specified events, which is more relevant to the target events and could precisely reveal the semantics of the event videos.

3. DISCOVERING CANDIDATE CONCEPTS FROM TAGS

In this section, we will present a three-step procedure to discover candidate concepts for each target event, which can be described in details as follows:

Step I: Flickr Image Crawling. Given a target event and its textual event description, we can use NLTK [4] to extract the nouns and verbs from the event definition sentence. Then we combine a noun and a verb to form a "noun-verb pair" as a textual query to perform text-based image search on Flickr. Finally, we downloaded the retrieved images and their associated tags for each query and combine them together as the concept discovery pool. Notably, the images retrieved in this way have higher relevance to the target event (See Figure 6).

Step II. Noisy Concept Filtering. Given a target event, we can crawl a number of images and their associated tags belonging to the same event from Flickr. As aforementioned, the tags are typically provided by general Internet users, and there are a fair amount of meaningless words that are irrelevant to the target event. To ensure that each tag corresponds to a meaningful concept, a tag filtering process is performed. Specifically, we use WordNet [24] as the concept lexicon and look up each tag in it. If a tag is matched successfully to a synset in WordNet, it is regarded as a meaningful concept. Otherwise, it is removed as a noisy word.

Step III. Concept Visualness Verification. After the filtering process, the remaining tags are meaningful concepts. Nevertheless, we notice that some concepts are not visually related. For example, there might be some images associated with concept "economy", but there are no consistent visual patterns within these images since the concept is highly abstract. Involving such concepts will bring significant distractions to video representation and degrade the final performance. Therefore, we need to verify the visualness of each concept so that only visually related concepts are included. To accomplish this, we first treat the images associated with a concept as positive training samples and simultaneously choose images from the other concepts as negative training samples. Then we split all training images into two halves and do a 2-fold cross validation. The performance is measured based on average precision. Finally, only the concepts with high cross validation performance are verified as visually related concepts and retained in the concept library. In order to obtain reliable concept detectors trained with sufficient number of images, we further remove the concepts that contain less than 80 training images. Noticed that each discovered concept is discovered for a specific event, we call the resultant concepts "event-specific concepts".

Based on the above strategy, we obtain the initial candidate concepts for each target event. Figure 2 shows the concept clouds discovered on two exemplary events, in which the size of the concept indicates its TF-IDF value, where we treat a set of tags in an event as a document.

4. BUILDING CONCEPT MODELS

In this section, we present how to choose reliable training images for each discovered concept, and then introduce how to build the corresponding concept model.

4.1 Training Image Selection for Each Discovered Concept

To get rid of the noisy and outlier images crawled from the Internet, we use a confidence ranking method to choose reliable training images. Given a concept c, we can construct the following two image subsets $\mathcal{X}^+ = \{\mathbf{x}_i\}_{i=1}^{m}$ and $\mathcal{X}^- = \{\mathbf{x}_i\}_{i=m+1}^{m+n}$, in which $\mathbf{x}_i \in \mathbb{R}^d$ is the d dimensional feature vector of the i-th image, $\mathcal{X}^+$ is a set of m images annotated with concept c, and $\mathcal{X}^-$ contains n images annotated without concept c. We adopt a soft neighbor assignment [9] in the feature space to estimate the confidence of assigning the given concept to an image. Particularly, each image $\mathbf{x}_i$ selects another image $\mathbf{x}_j$ as its neighbor with probability $p(\mathbf{x}_i, \mathbf{x}_j)$ and inherits its label from the image it selects. We define the probability $p(\mathbf{x}_i, \mathbf{x}_j)$ using a softmax operator over the entire image set $\mathcal{X} = \{\mathcal{X}^+, \mathcal{X}^-\}$:

$$p(\mathbf{x}_i, \mathbf{x}_j) = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2)}{\sum_{\mathbf{x}_k \in \mathcal{X} \setminus \{\mathbf{x}_i\}} \exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2)}, \quad (1)$$

where $\|\cdot\|$ denotes the L^2 norm of a vector.

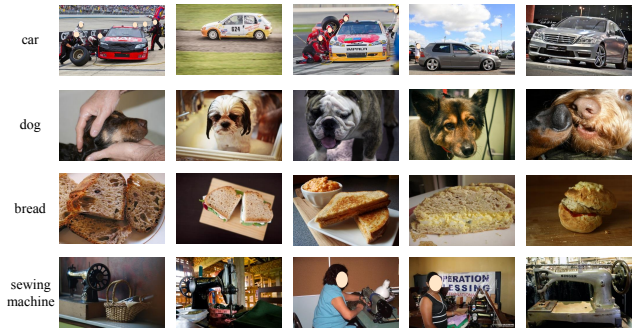


Figure 3: The top 5 images ranked by our method for some exemplary concepts.

Based upon this stochastic selection rule, we can calculate the probability $p(\mathbf{x}_i)$ that image $\mathbf{x}_i$ will be classified as positive with respect to the given concept:

$$p(c|\mathbf{x}_i) = \sum_{\mathbf{x}_j \in \mathcal{X}^+} p(\mathbf{x}_i, \mathbf{x}_j). \quad (2)$$

In the above equation, the confidence score of a concept's presence in an image is measured based on its visual closeness with respect to other images in the same concept category. If an image has noisy label, it tends to fall apart from the coherent visual pattern of the concept, leading to a small confidence value. Contrarily, images with correct labels will always comply with the common visual pattern of the concept, and thus are assigned with high confidence values. We use $p(c|\mathbf{x}_i)$ to estimate the confidence of image $\mathbf{x}_i$ belonging to concept c and select s images with the highest confidence scores as the positive training images for each concept. In this work, we set $s = 200$[1] and choose $t = 1,900$ negative images from other concepts as the negative training images. Figure 3 shows the top 5 images ranked by our method for some exemplary concepts. As can be seen, the selected images are highly relevant to the concepts while maintaining reasonable content diversity.

4.2 Concept Model Training

Given a concept c discovered for an event, suppose we have an Internet image collection $\mathcal{I} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{s+t}$, in which the label $y_i \in \{-1, 1\}$ of each image $\mathbf{x}_i$ is determined by the confidence score ranking method described in Section 4.1. In this work, we choose SVM classifier with RBF kernel as our concept model, in which the kernel function is defined as $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp\big(-d^2(\mathbf{x}_i, \mathbf{x}_j)/\sigma^2\big)$. Here $d(\mathbf{x}_i, \mathbf{x}_j)$ denotes the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_j$, and σ is the mean distance among all images on the training image set. We use LibSVM library [5] as the implementation of our SVM concept model and the optimal tradeoff parameter for SVM is determined via cross-validation.

4.3 Time Complexity Analysis

The overall complexity of concept model training consists of two parts: training image selection and concept model training. Specifically, the time complexity on choosing the training images described in Section 4.1 is $\mathcal{O}\big(d(m+n)^2\big)$, where m and n are respectively the number of positive and negative images for each concept and d is the feature dimension. In our experiment implemented on a MATLAB platform on an Intel XeonX5660 workstation with 3.2 GHz CPU and 18 GB memory, it takes 13.58 seconds to finish the confidence score calculation when $m = 3000$, $n = 3000$ and $d = 2,659$. On the other hand, the time complexity on concept model training described in Section 4.2 is $\mathcal{O}\big(d(s+t)^2 + (s+t)^3\big)$ which includes operations to compute the kernel and perform matrix inversion [5]. In our experiment with $s = 200$, $t = 1,900$ and $d = 2,659$, we finish the the training process of a concept model within 2 minutes on average. Considering the efficiency of the concept modeling process, our approach is applicable for constructing a large-scale concept library consisting of a huge number of concept models.

[1]If the images annotated with a concept are less than 200, we directly utilize all images as positive samples for concept modeling.

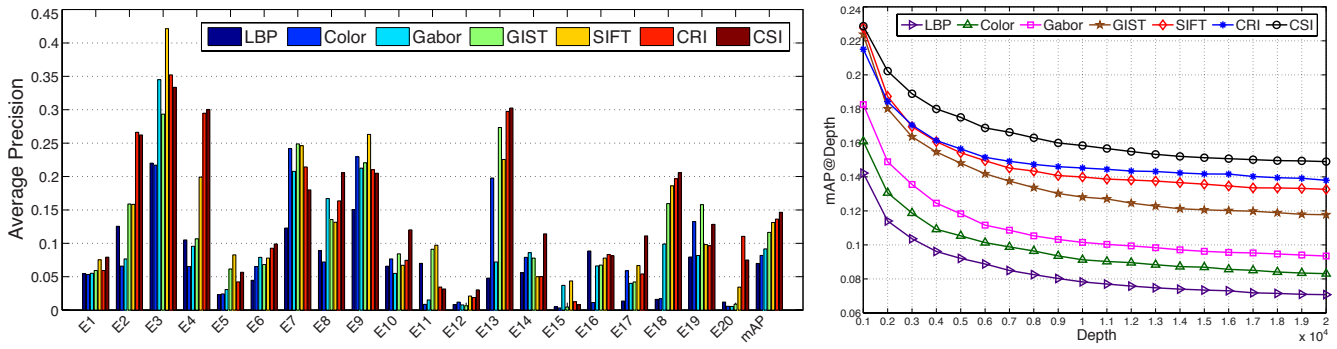
5. VIDEO EVENT DETECTION WITH DISCOVERED CONCEPTS

After constructing the models for all concepts in an event, we apply them on videos and adopt their probabilistic outputs as the concept-based representations, which can be used as an effective representation for semantic event analysis. In more details, given a video clip in a target event, we can first generate the concept based representation on each video frames and then average them as the final concept representation of the video clip, or we can directly apply the concept models on the averaged feature of the frames in the video. The second approach significantly reduces the concept score generation time and thus is adopted as the concept based video representation generation method in this work. There are typically two scenarios to apply concept based video representations for complex event detection, which can be discussed as below:

Scenario I: Supervised Event Modeling Over Concept Space. In this scenario, there are usually a number of labeled positive and negative training videos associated with a pre-specified event, and we regard the concept based video representation as high-level video content descriptors in the concept space for training a classifier of the target event. Therefore, we expect the concept based video representation to be discriminative so that the target event can be easily separated from other events. Given a pre-specified event detection task consisting of E events, we choose S concepts with the highest TF-IDF values for each of the E events from their respective discovered concepts, and concatenate the concept scores into an $E \times S$ dimensional feature vector (In this work, we set $E = 20$ and $S = 100$ and generate a $2,000$-dimensional concept based video representation for this task). With the concept feature representation as input, we can train any supervised model as the event classifier. In this work, we choose binary SVM classifier with χ^2 kernel as our event detection model. In the test stage, we adopt SVM probabilistic output as the event detection score on each test video, through which the video retrieval list can be generated.

Scenario II: Zero-Shot Event Retrieval. In this scenario, we do not have any training videos of the target event, but only directly use the event query[2] to retrieve relevant videos from the large video archive. Under this setting, the only available information is the concept scores on the test videos. We term this task zero-shot event retrieval since the procedure is purely semantic based. Since each concept has different levels of semantic relevance with respect to the query event, we use a weighted summation strategy to calculate the detection score of each test video. Given an event query e comprising of multiple words, we use WordNet [24], a large lexical database of English words, to estimate the semantic similarity of two words. The semantic relevance $r(e, c)$ between event e and concept c is determined as the maximum semantic similarity between concept c and all words appearing in event query e. We use a Python API for WordNet in NLTK [4] to calculate the Wu-Palmer Similarity [36] of two words. For an event and a test video with concept representation $\{s_1, \ldots, s_T\}$, in which each concept

[2]Event name and keywords extracted from event definition.



(a) AP on each event (b) mAP at variant depths

Figure 4: Performance of different methods on supervised event modeling task. CC: Classemes, CIN: ImageNet, CRI: proposed method without training image selection, CSI: proposed method with training image selection. This figure is best viewed in color.

score s_i corresponds to the event-specific concept c_i, the detection score of this video with respect to the target event can be estimated as $\sum_{i=1}^{T} r(e, c_i)s_i$. Finally, the event retrieval result can be generated by ranking the videos with these weighted summation scores.

6. EXPERIMENT

Our experiment aims to verify the effectiveness of our discovered concepts over the complex video event detection dataset. We begin with a description of the dataset, and then perform experiments evaluating different aspects of our method.

6.1 Dataset and Feature Extraction

Event Video Set. The TRECVID MED 2013 pre-specified task consists of a collection of Internet videos collected by Linguistic Data Consortium from various Internet video hosting sites [1]. The dataset contains $32,744$ videos falling into 20 event categories and the background category. The names of these pre-specified 20 events are respectively "E1: *birthday party*", "E2: *changing a vehicle tire*", "E3: *flash mob gathering*", "E4: *getting a vehicle unstuck*", "E5: *grooming an animal*", "E6: *making a sandwich*", "E7: *parade*", "E8: *parkour*", "E9: *repair an appliance*", "E10: *working on a sewing project*", "E11: *attempting a bike trick*", "E12: *cleaning an appliance*", "E13: *dog show*", "E14: *giving directions to a location*", "E15: *marriage proposal*", "E16: *renovating a home*", "E17: *rock climbing*", "E18: *town hall meeting*", "E19: *winning a race without a vehicle*", "E20: *working on a metal crafts project*". On this pre-specified event detection task in TRECVID MED 2013, each event is associated with a textual event kit that specifies the key concepts and detailed process of this event.

Internet Image Set. To collect Internet images, we utilize query words related to each event to perform image search on Flickr. For each event, we extract the keywords in the event textual kit and then try different combinations of any two keywords (typically one noun and one verb) to perform keyword based image search on *Flickr.com*, in which we combine all images from different queries together as the Internet images for this event. In this way, we downloaded $20,000$ images and their associated tags as the pool for concept discovery in each event. Specifically, we perform candidate concept discovery described in Section 3 and choose 100 potential concepts from the crawled image set for each event.

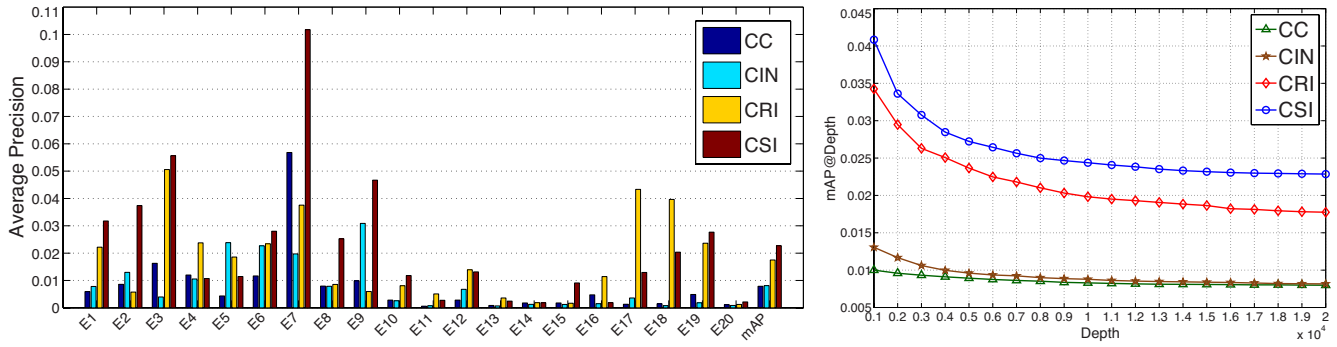
Feature Extraction. We extract $2,659$ dimensional classemes feature [32] as the feature representation of both Internet images and video frames. Given a video clip, we simply aggregate all frames of Classemes feature as the video-level feature representation. Other codebook based features such as SIFT BoW can also be used as the alternatives in our work.

6.2 Supervised Event Modeling Over Concept Space (MED EK100)

In this task, we treat the concept based video representations as feature descriptors for supervised event modeling. We follow the pre-defined training ($7,787$ videos) and test ($24,957$ videos) data split in the pre-specified EK100 task in TRECVID MED 2013 [1] in our experiment. There are 100 positive training videos and approximately 50 negative training videos in each event category. Moreover, both training and test sets contain significant number of background videos that do not belong to any target category, making the detection task very challenging. On each event, the Average Precision (AP), which approximates the area under the precision/recall curve, is adopted as evaluation metric of event detection. Finally, we further calculate mean Average Precision (mAP) across all 20 events as the overall evaluation metric on the entire dataset.

To evaluate the effectiveness of our discovered concept features in supervised event modeling, we compare the following middle or low level feature representations: (1) **SIFT** [19] Bag-of-Word (BoW). (2) **GIST** [27] BoW. (3) **Gabor** [8] BoW. (4) **LBP** [26] BoW. (5) Transformed **Color** Distribution [33] BoW. All the above five descriptors are densely extracted on grids of 20×20 pixels with 50% of overlap from images. For each type of the extracted descriptors, we train a codebook with 400 codewords, and partition each image into 1×1 and 2×2 blocks for spatial pyramid matching [15]. Finally, we adopt soft quantization [34] to represent each image as a $2,000$-dimensional histogram, which has the same dimension as our concept based video representation and ensures a fair comparison. (6) Concepts learned from Random Images (**CRI**). Concept models are learned using a randomly chosen subset of images associated with the concept tag directly without content consistency filtering as described in Section 4.1. (7) Our proposed $2,000$ Concepts learned from Selected Images (**CSI**). We use our method to select reliable training images for concept modeling.

Figure 4(a) illustrates the performance of all the methods on this task quantitatively. From the results, we can have the following observations: (1) The concepts generated by our CSI method consistently beats others by a large margin, which demonstrates its effectiveness in concept based video representation. On some events where our method is inferior to other baselines, we observe significant domain difference between Flickr images and MED videos. We will consider this issue by exploring cross domain adaptation methods in our future work. (2) The CSI method beats the five types of low-level features, which implies that our discovered concepts can not only reveal the semantic concepts but also be utilized as an effective feature description for event discrimination. (3) The CSI method performs significantly better than the CRI method.



(a) AP on each event (b) mAP at variant depths

Figure 5: Performance of different concept-based classifiers for zero-shot event detection task. This figure is best viewed in color.

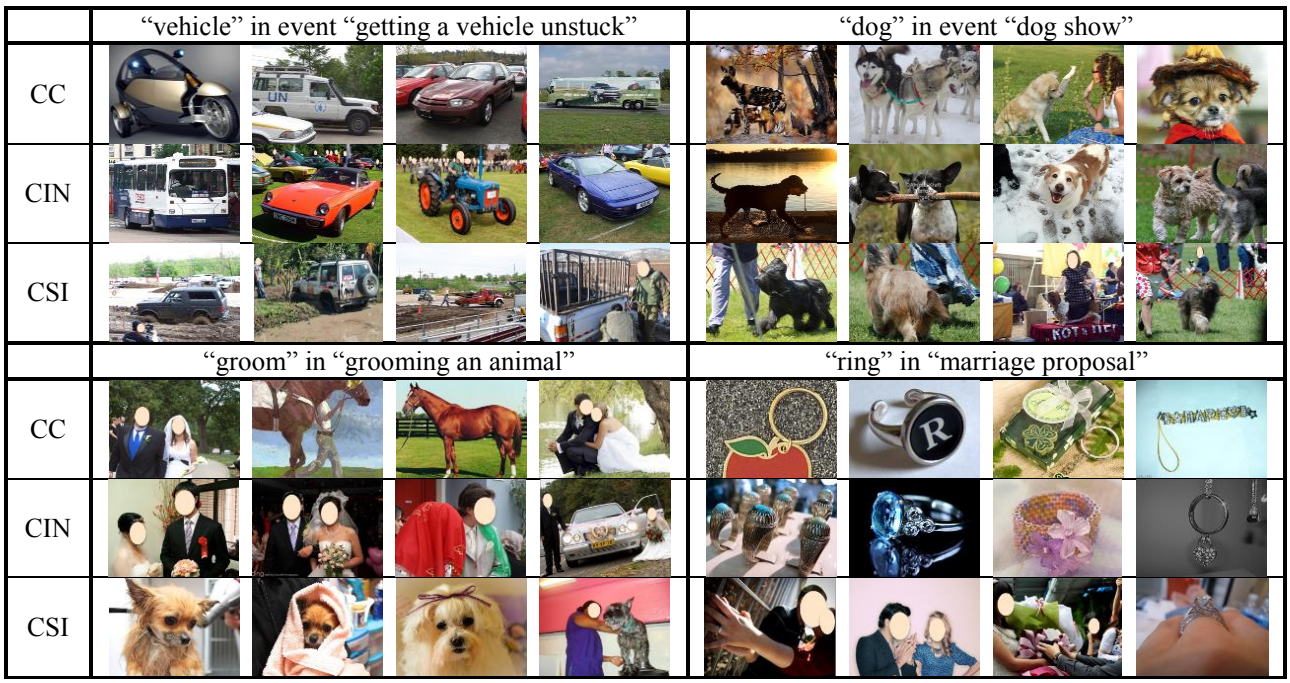


Figure 6: Concept training images form different concept sources. Note that the training images utilized in our method (CSI) not only contain the concept but also convey context information about the event.

This is due to the fact that the former leverages more reliable training images than the random images utilized in the latter. This verifies the soundness of our confidence ranking method in selecting clean training images for concept modeling. Figure 4(b) shows the mAP comparisons at variant returned depths (i.e., the number of top ranked test videos included in the evaluation). From the results, we can see that our method achieves significant and consistent mAP improvements over the other methods at variant returned depths.

6.3 Zero-Shot Event Retrieval (MED EK0)

In this task, we directly apply the concept scores to rank the videos in the archive without leveraging any training video sample. Similar to Section 6.2, we also adopt AP and mAP as our evaluation metric. The focus of this experiment is to reveal the effectiveness of the discovered semantic concepts compared with semantic concepts discovered from other methods. To this end, the following concepts generated from different methods will be compared: (1) Classemes Concepts (**CC**). We extract the $2,659$-dimensional classemes concept feature from the videos. Given each event query, we use the WordNet Wu-Palmer semantic similarity between event query and the concept names (see Section 5) in Classemes to find 100 most relevant Classemes concepts for this event. (2) Concepts discovered from ImageNet (**CIN**) [6]. In this method, we want to discover concepts for each event from all the concepts in ImageNet. For each event, we also adopt WordNet to calculate the semantic similarity between its event query and the concept names in ImageNet, and choose the same number of concepts (100 for each event) from ImageNet . For each concept, we choose 200 images from its ImageNet synset as the positive training images and $1,900$ images from other discovered ImageNet concepts as the negative training images. These images are then fed into a SVM classifier as the concept model. (3) Concepts learned from Random Images (**CRI**), in which we randomly pick up the same number of images as used in our method as training images for concept modeling. (4) Our proposed Concepts learned from Selected Images (**CSI**).

Figure 5(a) shows the per-event performance of all the methods. In Figure 5(b), we further plot the mAP at different returned depths for different comparison methods. From the results, we have the following observations: (1) our CSI method achieves the best performance (with a relative performance gain as high as 228%) over

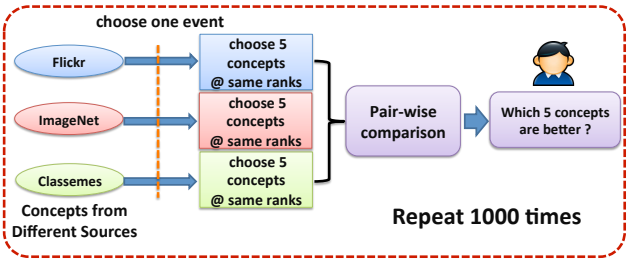


Figure 7: Human evaluation of concepts discovered from different sources (Flickr, ImageNet, and Classemes).

most of the events. Since the task is purely semantic based, the results clearly verify that the concepts discovered by our method are applicable for semantic based event retrieval. (2) The zero-shot event retrieval performance is worse than that of supervised event modeling. This is because that the latter uses training videos to obtain a more sophisticated event model while the former is merely based on concept score aggregation. (3) Our CSI method performs much better than Classemes, since most concepts in Classemes are irrelevant to the events. (4) The CSI method clearly beats the CIN method. This shows that our concept discovery method is able to obtain more accurate concept representations to the events than the concept representations discovered from other lexicon such as ImageNet. The reason may be two-fold: First, the concepts discovered from Flickr images are more relevant to the real-world events than the general concepts in ImageNet. In fact, our discovered concepts are the most semantically relevant to the events comparing with the concepts discovered from Classemes and ImageNet, as illustrated in Table 1. Second, the Flickr images used in training our concept model contain visual clues of the event, while the images in ImageNet usually contain clean objects without event background. Figure 6 illustrates the training images from Flickr and ImageNet for some concepts. We can see that for the concept "groom" in event "grooming an animal", because of the ambiguity of the word "groom", training images from CC and CIN generally refer to "bridegroom" while our concept refers to the action of "grooming an animal". Another example is the concept "ring" in event "marriage proposal", where training images from CC and CIN are general "ring" with simple background and do not contain any context information about marriage proposal.

6.4 Human Evaluation of Concepts

Besides the performance comparison described in the previous sections, we also design an experiment to ask human subjects to evaluate the quality of concepts discovered from different sources including Classemes, ImageNet and our proposed framework using Flickr images. Details of the experiment are illustrated in Figure 7.

For each evaluation session, we first randomly pick up one event from the 20 pre-specified events in TRECVID MED 2013, and generate 100 concepts from each of the three sources (Flickr, ImageNet, Classemes) using the approach described in Section 6.3. In this step, we rank the 100 concepts from each source in descending order based on the WordNet semantic similarity between concept name and event query. Second, from each of the three ranked concept lists, we randomly choose 5 concepts at the same rank positions, and form three concept subgroups, each of which contains the sampled 5 concepts. Third, we randomly pick up two concept subgroups from the three to form a pair and ask a user to judge which subgroup of 5 concepts is more relevant to the event.

The above procedure is repeated $1,000$ times, through which we obtain the following statistics about concept quality. (1) Flickr is better than the other two sources with 81.29% chance. (2) ImageNet is better than others with 34.19% chance. (3) Classemes is better than others with a 32.46% chance. From these results, we can see that our discovered concepts from Flickr images are more relevant to the events based on subjective evaluation by humans.

Figure 8: Semantic recounting for videos from exemplary events in TRECVID MED 2013: each row shows evenly subsampled frames of an example video and the top 5 relevant concepts detected in the video.

6.5 Video Semantic Recounting

For each video of a target event, we rank all the concepts discovered for the event based on their confidence scores and treat the top ranked concepts as the semantic description of the video content. Such a procedure is able to reveal the semantic information contained in a video and is thus called video semantic recounting. Figure 8 shows the recounting result on videos from some exemplary events in TRECVID MED 2013, in which the top 5 ranked concepts generated by our method are selected as concepts for each video. As can be seen, these concepts reveal the semantics contained in the videos, which verifies the effectiveness of the our discovered concepts in representing video semantics.

7. CONCLUSION

We have introduced an automatic event driven concept discovery method for semantic based video event detection. Given a target event, we crawl a collection of Flickr images and their associated tags related to this event as the concept discovery knowledge pool. Our method first estimate the candidate concepts present in the event by measuring the visualness of each concept and its semantic meaning. Then a concept model is constructed for each candidate concept based on the large margin SVM classifier. Finally, the individual concept models are applied on the event videos to generate concept-based video representations. We test our discovered concepts over two video event detection tasks including supervised event modeling over concept space and zero-shot event retrieval, and the promising experiment results have demonstrated the effectiveness of the proposed event-driven concept discovery method. For future work, we will explore the spatial-temporal concept sequences in dynamic videos and investigate their effectiveness in complex event detection.

Event Name	Concepts Discovered from Different Sources	
getting a vehicle unstuck	Classemes	air transportation vehicle, all terrain vehicle, amphibious vehicle, armed person, armored fighting vehicle, armored recovery vehicle, armored vehicle, armored vehicle heavy, armored vehicle light, command vehicle
	ImageNet	vehicle, bumper car, craft, military vehicle, rocket, skibob, sled, steamroller, wheeled vehicle, conveyance
	Ours	**tire, car, snow, stick, stuck, winter, vehicle, truck, night, blizzard**
grooming an animal	Classemes	adult animal, animal, animal activity, animal blo, animal body part, animal body region, animal cage, animal container, animal pen, animal shelter
	ImageNet	groom, animal, invertebrate, homeotherm, work animal, darter, range animal, creepy-crawly, domestic animal, molter
	Ours	**dog, pet, grooming, cat, animal, bath, cute, canine, puppy, water**
making a sandwich	Classemes	baking dish, cafe place, classroom setting, collection display setting, cutting device, dish drying rack, food utensil, hair cutting razor, hdtv set, hole making tool
	ImageNet	sandwich, open-face sandwich, butty, reuben, ham sandwich, gyro, chicken sandwich, hotdog, club sandwich, wrap
	Ours	**sandwich, food, bread, cooking, cheese, spice, baking, pan, kitchen, breakfast**
working on a sewing project	Classemes	clothes iron, landing craft, laundry room, living room, missile armed craft, multi room unit, work environment, work station, steel mill worker, carpentry tool
	ImageNet	sport, outdoor game, rowing, funambulism, judo, blood sport, gymnastics, water sport, track and field, outdoor sport
	Ours	**sewing, handmade, embroidery, craft, quilt, fabric, hand, sewing machine, textile, thread**
cleaning an appliance	Classemes	action on object, animal container, armed person, art object, back yard, bag, bilateral object, box the container, butcher shop, capsule container
	ImageNet	appliance, gadgetry, gimbal, injector, mod con, device, musical instrument, acoustic device, adapter, afterburn
	Ours	**kitchen, furniture, washing, bed, sink, divan, spring bed, cleaning, stove, dishwasher**
rock climbing	Classemes	astronomical observatory building, attached body part, auto part, bar building, body movement event, body of water, building, building cluster, building security system, cavity with walls
	ImageNet	rock, uphill, outcrop, whinstone, xenolith, tor, slope, ptyalith, kidney stone, urolith
	Ours	**climbing, rock climbing, bouldering, mountain, sport, hiking, climber, landscape, peak, rope**

Table 1: Top 10 **concepts discovered form different sources.**

8. ACKNOWLEDGEMENT

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

9. REFERENCES

[1] http://www.nist.gov/itl/iad/mig/med13.cfm.

[2] S. Ayache and G. Quénot, Video Corpus Annotation using Active Learning. In *ECIR*, 2008.

[3] T. Berg, A. Berg, and J. Shih. Automatic Attribute Discovery and Characterization from Noisy Web Data. In *ECCV*, 2010.

[4] S. Bird. NLTK: The Natural Language Toolkit. In *ACL*, 2006.

[5] C.-C. Chang and C.-J. Lin. LIBSVM : A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, ImageNet: A large-scale Hierarchical Image Database. In *CVPR*, 2009.

[7] L. Duan, D. Xu, I. W. Tsang, and J. Luo, Visual Event Recognition in Videos by Learning from Web Data. In *CVPR*, 2010.

[8] D. Field, et al. Relations between the statistics of natural images and the response properties of cortical cells *J. Opt. Soc. Am. A*, 1987.

[9] J. Goldberger, S. Roweis, G. Hinton, and R. Salakhutdinov. Neighbourhood Components Analysis. In *NIPS*, 2004.

[10] N. Ikizler-Cinbis, R. Cinbis, and S. Sclaroff, Learning Actions from the Web. In *ICCV*, 2009.

[11] H. Izadinia, and M. Shah. Recognizing Complex Events using Large Margin Joint Low-Level Event Model. In *ECCV*, 2012.

[12] Y.-G. Jiang, S. Bhattacharya, S.-F. Chang, M. Shah. High-Level Event Recognition in Unconstrained Videos. *IJMIR*, 2013.

[13] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. Columbia-UCF TRECVID2010 Multimedia Event Detection: Combining Multiple Modalities, Contextual Concepts, and Temporal Matching. In *NIST TRECVID Workshop*, 2010.

[14] L. Laptev, M. Marszaek, C. Schmid, and B. Rozenfeld. Learning Realistic Human Actions from Movies. In *CVPR*, 2008.

[15] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[16] L.-J. Li, H. Su, L. Fei-Fei, and E. Xing. Object Bank: A high-level Image Representation for Scene Classification & Semantic Feature Sparsification. In *NIPS*, 2010.

[17] J. Liu and M. Shah. Learning Human Actions via Information Maximization. In *CVPR*, 2008.

[18] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. Sawhney. Video Event Recognition Using Concept Attributes. In *WACV*, 2012.

[19] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[20] Z. Ma, Y. Yang, Y. Cai, N. Sebe, and A. Hauptmann. Knowledge Adaptation for Ad Hoc Multimedia Event Detection with Few Exemplars. In *ACM MM*, 2013.

[21] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. Hauptmann. Complex Event Detection via Multi-Source Video Attributes. In *CVPR*, 2013.

[22] M. Mazloom, E. Gavves, K. Sande, and C. Snoek. Searching Informative Concept Banks for Video Event Detection. In *ICMR*, 2013.

[23] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic Model Vectors for Complex Video Event Recognition. *IEEE TMM*, 2012.

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

[25] P. Natarajan, S. Wu, S. Vitaladevuni, and X. Zhuang. Multimodal Feature Fusion for Robust Event Detection in Web Videos. In *CVPR*, 2012.

[26] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *TPAMI*, 2002.

[27] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 2001.

[28] J. Revaud, M. Douze, C. Schmid, and H. Jégou. Event Retrieval in Large Video Collections with Circulant Temporal Encoding. *CVPR*, 2013.

[29] S. Sadanand and J. Corso. Action Bank: A High-level Representation of Activity in Video. In *CVPR*, 2012.

[30] A. Tamrakar, S. Ali, Q. Yu, J. Liu, O. Javed, A. Divakaran, H. Cheng, and H. Sawhney. Evaluation of Low-level Features and Their Combinations for Complex Event Detection in Open Source Videos. *CVPR*, 2012.

[31] K. Tang, L. Fei-Fei, and D. Koller. Learning Latent Temporal Structure for Complex Event Detection. In *CVPR*, 2012.

[32] L. Torresani, M. Szummer, and A. Fitzgibbon. Efficient Object Category Recognition Using Classemes. In *ECCV*, 2010.

[33] K. Van De Sande, T. Gevers, and C. Snoek. Evaluating color descriptors for object and scene recognition. *TPAMI*, 2010.

[34] J. van Gemert, C. Veenman, A. Smeulders, and J-M Geusebroek. Visual word ambiguity *TPAMI*, 2010.

[35] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *CVPR*, 2011.

[36] Z. Wu and M. Palmer. Verbs Semantics and Lexical Selection. In *ACL*, 1994.

[37] K. Yanai and K. Barnard. Image Region Entropy: A Measure of "Visualness" of Web Images Associated with One Concept. In *ACM MM*, 2005.

[38] Y. Yang and M. Shah. Complex Events Detection using Data-driven Concepts. In *ECCV*, 2012.