Composite Concept Discovery for Zero-Shot Video Event Detection

Amirhossein Habibian, Thomas Mensink, and Cees G. M. Snoek
ISLA, Informatics Institute, University of Amsterdam
Science Park 904, 1098 XH, Amsterdam, The Netherlands
{a.habibian, thomas.mensink, cgmsnoek}@uva.nl

ABSTRACT

We consider automated detection of events in video without the use of any visual training examples. A common approach is to represent videos as classification scores obtained from a vocabulary of pre-trained concept classifiers. Where others construct the vocabulary by training individual concept classifiers, we propose to train classifiers for combination of concepts composed by Boolean logic operators. We call these concept combinations *composite concepts* and contribute an algorithm that automatically discovers them from existing video-level concept annotations. We discover composite concepts by jointly optimizing the accuracy of concept classifiers *and* their effectiveness for detecting events. We demonstrate that by combining concepts into composite concepts, we can train more accurate classifiers for the concept vocabulary, which leads to improved zero-shot event detection. Moreover, we demonstrate that by using different logic operators, namely "AND", "OR", we discover different types of composite concepts, which are complementary for zero-shot event detection. We perform a search for 20 events in 41K web videos from two test sets of the challenging TRECVID Multimedia Event Detection 2013 corpus. The experiments demonstrate the superior performance of the discovered composite concepts, compared to present-day alternatives, for zero-shot event detection.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

Keywords

Event recognition, Concept representation

1. INTRODUCTION

We address the problem of zero-shot event detection in video, where the goal is to detect complex events without the use of *any* visual training examples. Since there is no visual example available to train event classifiers, the common approach in zero-shot detection literature relies on textual specification of events to extract the event models. For this purpose, each event is modeled by specifying its relevant and irrelevant semantic concepts, which are identified from text. Afterward, test videos are ranked based on their similarity to the event model. For this purpose, every test video should be represented by its semantic features, such as transcripts from automatic speech recognition (ASR) [14, 3], text from video optical character recognition (VOCR) [14], and scores from concept classifiers [22, 3]. Using ASR and VOCR features leads to high precision in event detection, as these sources generally result in reliable textual descriptions. However, not all videos come with ASR and VOCR features, leading to low recall in event detection. In contrast, high recall is reported for zero-shot detection using concept classifiers, but the low accuracy of concept classifiers leads to poor precision in detecting events [3]. In this paper, we consider zero-shot event detection using concept-classifiers.

Figure 1: Video examples for training vocabulary concept classifiers accompanied with their concept annotations. Positive and negative labels are denoted by 1 and 0. Primitive and composite concept annotations are shown in black and red colors. Composite concepts annotations are inferred from their primitive concepts.

ICMR'14, April 1–4, 2014, Glasgow, United Kingdom.

In zero-shot event detection by concept classifiers, videos are represented as the output of a vocabulary of concept classifiers. The common approach is to train a one-against-all classifier per concept in the vocabulary [22, 3, 13, 5]. Hence, the underlying assumption for these approaches is that the concepts in the vocabulary, and their annotations, are independent from each other. This strong assumption may not always be valid. Consider for example the concept classifiers "Bike" and "Motorcycle", which are very likely to have a considerable overlap in the visual context in which they may appear. Consequently, we argue it is advantageous to exploit the visual (and semantic) consistency of various concept classifiers. Our intuition is that there are combinations of concepts, for which training one joint-classifier is more effective than separately training concept classifiers. Continuing the example, combining the positive annotations for "Bike" and "Motorcycle" into a single concept classifier may result in a more reliable classifier for zero-shot event detection. In addition, some of the concept combinations may be very descriptive and characteristic for an event, *i.e.,* the combination of "Bike" and "Ride" better characterizes the event "attempting bike trick" than the "Bike" and "Ride" concepts individually. In this paper, we consider the interrelation between concepts before training vocabulary concept classifiers. Based on concept relations, we combine vocabulary concepts for each event so as to optimize the event detection accuracy.

Others have also investigated optimizing the vocabulary concepts for event detection [3, 11, 15]. In [11], for example, an iterative feature selection algorithm is proposed that learns from examples to select a subset of pre-trained concept classifiers that optimizes event detection accuracy. We also aim for optimizing the vocabulary concepts per event, but rather than selecting from rigid concept classifiers only, we introduce a flexible composition of classifiers by adapting their underlying training data. This bears similarity to recently introduced bi-concepts [8] and visual phrases [4] for concept detection. In these works, the concept combination is defined as the co-occurrence of annotations in training data. Learning concept detectors from these co-occurred annotations results in more effective and descriptive classifiers. We observe that bi-concepts combine concept annotations by logical "AND" relations. In this paper, we generalize this combination logic by also considering other logical relationships in particular logical "OR", *i.e.,* "Bike-OR-Motorcycle". We call these logical combinations *composite concepts* and define them as the logical composition of *primitive concepts*, see Figure 1.

The main challenge in constructing composite concepts is to discover which primitive concepts should be combined together. This problem has been studied by Rastegari *et al.* [19], for discovering bi-concepts for image search. They discover bi-concepts by searching for concept pairs whose joint classifier is more accurate than individual concept classifiers. However, combining the concepts by only considering their classification accuracy might fail for event detection. For example, combining the concepts "Dog" and "Cat" might lead to a more accurate concept classifier. However, the "Dog" concept is individually more effective for detecting the event *dog show* and it loses its effectiveness when combined with other concepts. Different from the references, we propose an algorithm that automatically discovers composite concepts by jointly considering the accuracy of concept classifiers *and* their effectiveness for detecting events.

The main contribution of this paper are: First, we propose the notion of composite concepts for constructing concept vocabularies for zero-shot event detection. Second, we propose an algorithm to automatically discover composite concepts from a vocabulary of primitive concepts. Third, our experiments on the challenging TRECVID Multimedia Event Detection 2013 corpus demonstrates the effectiveness of composite concepts for zero-shot event detection.

2. RELATED WORK

Representing videos as scores from a vocabulary of concept classifiers is shown to be promising for event detection [10, 13, 5, 7], especially when only few [12, 22] or no visual example of the events [3, 9] are available. In the references, the vocabulary concept classifiers are trained from a set of images or videos, which are annotated with respect to presence or absence of the concepts. The common approach is to train a separate classifier per concept. Different from the references, we train concept classifiers from combinations of vocabulary concept annotations, as composite concepts.

Optimizing the vocabulary concepts per event has recently attracted research attention. In [3, 11], for example, the vocabulary is optimized for each event by automatically selecting a subset of concepts per event. In [15] the Wordnet ontology is used to measure the relevance between each query and the available vocabulary concepts, in order to select the most relevant concepts. Dalton *et al.* [3] select the relevant concepts per event by considering concept dependencies modeled by a Markov Random Field. Mazloom *et al.* [11] rely on supervised feature selection to select the most discriminative concepts per event. Our paper is different from these works in the following two ways. First, sometimes two concepts are individually uninformative for an event but their composition is informative and should be kept in the vocabulary. Therefore, instead of excluding the uninformative concepts from the vocabulary, we search for a composition where the concepts are informative. Second, in addition to considering the informativeness of concepts for events, we also consider concept classification accuracy. Therefore, we include the concepts in the vocabulary which are not only informative for an event, but are also accurately classified.

Rather than combining the concept classifiers in the vocabulary a posteriori, one can also combine them a priori by considering concept interrelationships during training concept classifiers [19, 4, 8, 2, 23]. In [4, 8] the notion of "bi-concepts" or "visual phrase" are introduced as the co-occurrence of distinct concepts which correspond to a very characteristic appearance that makes their detection, as one concept, more effective. Bi-concepts have been used for various purposes: sentiment analysis [2], where the sentiment concept classifiers are trained as bi-concepts of adjective and nouns, *i.e.,* cute dog. Image search [8], where co-occurring query concepts are trained together as bi-concepts. Object recognition [4], where co-occurring objects are trained as one bi-concept. Moreover, Rastegari *et al.* [19] proposes to automatically discover bi-concepts from query concepts for image search. They consider a combination of concepts as a bi-concept only if it is more effectively classified than its individual concepts.

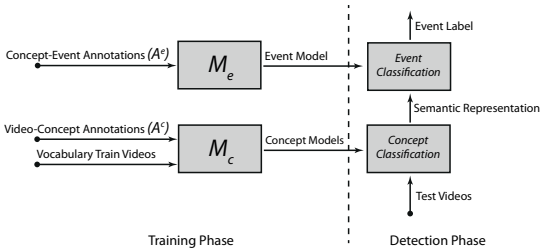

Figure 2: Data flow diagram of our defined zero-shot event detection pipeline. The notations are explained in Section 3.1.

Our paper is different from these works in the following ways. First, bi-concepts can be considered as one type of all possible concept compositions, where concepts are composed by "AND" relations. Differently, our proposed composite concepts can be discovered from any type of concept relations, including "OR", "AND", "XOR" etc. Second, Rastegari *et al.* [19] discover bi-concepts only based on their classification accuracy and ignore their effectiveness when classifying events. Differently, we discover composite concepts by jointly considering their classification accuracy *and* their effectiveness for recognizing events. As we will show in the experiments, this criteria has a high impact on performance of the composite concepts in event detection.

3. COMPOSITE CONCEPT DISCOVERY

We first formalize our zero-shot event detection settings. Then we propose our algorithm for discovering composite-concepts. Finally, we discuss about the computational costs of the proposed algorithm.

3.1 Zero-Shot Event Detection

In zero-shot event detection the goal is to detect an event, without using any visual examples of the event, see Figure 2 for an illustration. The detection problem is decomposed into two parts, first, low-level video features are mapped into an intermediate semantic representation, and second, the semantic representation is mapped to an event. In this work, we consider a set of concepts $C = \{c_1, \ldots, c_l\}$ as the intermediate representation, and they cover a wide range of concepts including objects, actions, and scene concepts, such as the concepts recommended in [5].

To obtain the intermediate representation, we learn a mapping, $\mathcal{M}_c : \boldsymbol{x} \mapsto \boldsymbol{a}$ from a d-dimensional video feature $\boldsymbol{x} \in \mathbb{R}^d$ to an ℓ-dimensional binary vector $\boldsymbol{a} \in \{0,1\}^\ell$, denoting the presence/absence of the concepts in C. $\mathcal{M}_c$ is learnt from a set of annotated videos, denoted as $\mathcal{A}^c = \{(\boldsymbol{x}_i, \boldsymbol{a}_i)\}_{i=1}^I$, which we call the *video-concept annotations*. Using these annotation the mapping is obtained by training a classifier for each concept in C independently.

For the event detection, we learn a mapping, $\mathcal{M}_e : \boldsymbol{a} \mapsto e$ from the intermediate representation $\boldsymbol{a}$ to the event label $e \in \{0,1\}$. In our work, $\mathcal{M}_e$ is obtained from a train set of textual descriptions of events, denoted by $\mathcal{A}^e = \{(\boldsymbol{a}_j, e_j)\}_{j=1}^J$, which we coin the *concept-event annotations*. This can be provided at category level or at instance level. For category level descriptions, there is just a single concept annotation per event. For instance level annotations, multiple concept-event annotations are provided for the same event denoting possibly a variety of concepts related to the event. We use instance level concept-event annotations, since these are usually easier to obtain[1]. Using the annotation set $\mathcal{A}^e$, mapping $\mathcal{M}_e$ is obtained by training an event classifier from concept annotations as features.

Note that, the two mappings $\mathcal{M}_c$ and $\mathcal{M}_e$ are learned from a different train set, and that for training $\mathcal{M}_e$ no visual example is used. The performance of such a zero-shot detection framework, critically depends on the quality of the chosen intermediate representation.

3.2 Algorithm

In this section we detail our algorithm to learn a set of composite concepts for the zero-shot prediction of an event. We propose to use composite concepts as intermediate representation which are learned for a specific event e. Composite concepts are derived from the set of primitive concepts C, as a combination of two (or more) primitive concepts, with a logical operator. For simplicity, we only describe the use of the logical OR operator, but the described algorithm holds for the other operators as well.

Finding the composite concepts is reducible to a set partitioning problem. The goal is to find a set of composite concepts $\widehat{C}$, which is a division of primitive concepts C, as a union of non-overlapping and non-empty subsets. For a set C of ℓ primitive concepts, the total number of possible partitions is the Bell number $B_\ell = \sum_{k=0}^{\ell-1} \binom{\ell-1}{k} B_k$. Finding an optimal $\widehat{C}$ is an NP-hard problem, see *e.g.* [19], for which we propose a greedy approximation.

The accuracy of the event detection accuracy depends on two competing objectives, *concept predictability* and *event predictability*. Concept predictability measures the accuracy of the prediction of the concepts from the intermediate representation. Event predictability measures the discriminative power of the intermediate representation to detect an event. Based on these two criteria, we find the set of composite concepts $\widehat{C}$ for event e by maximizing:

$$S_e(\widehat{C}\,) = \lambda P_c(\widehat{C}\,) + (1-\lambda) P_e(\widehat{C}\,) \tag{1}$$

where P_c and P_e measure the concept predictability and event predictability, respectively, and will be described in detail below. Moreover, $\lambda \in [0,1]$ is a parameter to balance between concept predictability and event predictability. This parameter can be optimized by cross-validation. However, when visual examples of events are not available for cross-validation, we pick $\lambda = 0.5$ to equally weight concept predictability and event predictability.

Our proposed greedy approximation is similar to hierarchical clustering. Starting from the set of primitive concepts $C_1 \leftarrow C$, in each iteration t we find two concepts n^* and m^* to be merged. The two concepts are selected based on the expected improvement of Eq. 1:

$$(m^*, n^*) = \underset{m,n \in C_t \text{s.t.} m \neq n}{\operatorname{argmax}} \Delta S_e(C_t, m, n), \tag{2}$$

where $\Delta S_e(C, m, n)$ denotes the difference between using C_t and the set where m and n are used as composite concept. We use n^* and m^* to define C_{t+1}, as:

$$C_{t+1} = C_t - C_{n^*} - C_{m^*} + \big(C_{n^*} \vee C_{m^*}\big). \tag{3}$$

[1]Generalizing to category level annotations is straight forward.

input : C, $\mathcal{A}^c$ and $\mathcal{A}^e$
output: $\widehat{C}$
$C_1 \leftarrow C$
for $t \leftarrow 1$ ***to*** ℓ **do**
  compute ΔS_e for each pair $m, n \in C$
  $(m^*, n^*) \leftarrow \text{argmax}\, \Delta S_e$
  if $\Delta S_e(m^*, n^*) > 0$ **then**
    $C_{t+1} = C_t - C_{n^*} - C_{m^*} + \big(C_{n^*} \vee C_{m^*}\big)$
  else
    return $\widehat{C} \leftarrow C_t$
  end
end

Algorithm 1: Pseudo code for the proposed algorithm to discover composite concepts for the event e. Notation conventions are detailed in Section 3.2.

The clustering algorithm is terminated at iteration t' when $\Delta S_e(C_t, m, n) < 0$, The final set of composite concepts $\widehat{C}$ used for this event is $\widehat{C} \leftarrow C_{t'}$. The Pseudo code of our clustering procedure is summarised in Algorithm 1.

3.2.1 Concept predictability

We measure the concept predictability of a set of concepts C, by their classification performance on a part of the train set. Therefore, we first learn the mapping $\mathcal{M}_c$ for each concept $c \in C$, using the annotation $\mathcal{A}^c$. Then we evaluate this mapping on a hold out partition of training data, using the average precision measure, denoted as $P_{\text{AP}}(c)$. So, the concept predictability is given by:

$$P_c(C) = \frac{1}{|C|} \sum_{c \in C} P_{\text{AP}}(c). \tag{4}$$

In contrast to [19], this notion of concept predictability relies on the accuracy of the classifiers, rather than an estimation based on a geometric intuition.

For the greedy algorithm, we are interested in relative improvement of C_t when merging concepts m and n. Let $k = |C_t|$, and $C_{t'}$ denote the set C_t minus c_m and c_n, then:

$$\begin{aligned}\Delta P_c(C_t, m, n) &= \frac{1}{k-1} \sum_{c \in C_{t'}} P_{\text{AP}}(c) + \frac{1}{k-1} P_{\text{AP}}(c_m \vee c_n) \\ &\quad - \frac{1}{k} \sum_{c \in C_{t'}} P_{\text{AP}}(c) - \frac{1}{k}\big(P_{\text{AP}}(c_m) + P_{\text{AP}}(c_n)\big) \\ &\approx P_{\text{AP}}(c_m \vee c_n) - \frac{1}{2}\big(P_{\text{AP}}(c_m) + P_{\text{AP}}(c_n)\big),\end{aligned} \tag{5}$$

for the approximation in the final step, we assume that $\sum_{c \in C_{t'}} P_{\text{AP}}(c)$ has a value independent of the chosen m and n. For each iteration of the clustering algorithm, we train a classifier for each pair of concepts m and n, and compute the value of $\Delta P_c(C_t, m, n)$ by using Eq. 5.

3.2.2 Event predictability

The event predictability is measured by the quality of the mapping from the composite concepts to the event label, using the train set $\mathcal{A}^e$. More precisely, given a set of composite concepts C_t, we train an event classifier on the concept annotations $\boldsymbol{a}_j$ as features, apply the composition given by C_t, and using the labels e_j as desired outcome. Then the trained classifier is evaluated on a hold out partition of training data using average precision. The event predictability is given by:

$$P_e(C) = P_{\text{AP}}(e), \tag{6}$$

where C denotes the composition which should be applied on the concept annotations $\boldsymbol{a}$.

For the greedy algorithm, we evaluate the relative improvement of merging concepts m and n of set C_t, by:

$$\Delta P_e(C_t, m, n) = P_e(C_t, m, n) - P_e(C_t). \tag{7}$$

3.3 Computational Efficiency

For a vocabulary of n primitive concepts our algorithm needs to train $\mathcal{O}(n^2)$ concept classifiers to discover composite concepts. In the first iteration $\frac{n(n-1)}{2}$ concept classifiers are trained. By storing the accuracies of the concept classifiers in the first iteration, we only have to train $n-1$ new concept classifiers, in the second iteration, to compare the accuracy of the new composite concept to the existing $n-1$ classifiers. Similarly $n-2$ concept classifiers are trained in the third iteration and so on, which leads to an overall computational complexity of $\mathcal{O}(n^2)$ in training concept classifiers.

Even though we use an efficient greedy approximation of the original learning problem, the computational complexity is still rather high. This is due to the high dimensional video features used for training the concept classifiers, as required by Eq. 5.

However, Eq. 5 shows that we are interested in the *performance differences* between concepts and composite concepts, not in the maximum achievable performance of these (composite) concepts. Visual classifiers, especially in a large-scale setting with many examples or concepts, can be effectively obtained by using stochastic gradient descent [18]. Stochastic gradient descent is an optimization algorithm, which is used to gradually train the classifiers by randomly passing through the training data. For good performance a high number of epochs are required, typically in the order of $100 - 500$, however, for the performance difference we could use the classifier obtained after just a few epochs, since this is typically a good predictor of the expected performance.

Therefore, we use very early stopping (after 5 epochs) and a large value of the learning rate parameter, to ensure rapid learning from the given examples.

4. EXPERIMENTAL SETUP

4.1 Datasets

Video Data: We perform our experiments on the challenging TRECVID Multimedia Event Detection 2013 corpus, containing in total 51K arbitrary videos collected from the web. To the best of our knowledge this is the largest publicly available video corpus in the literature for event detection containing user-generated video with a large variation in quality, length and content. We perform our experiments on three partitions of videos: Research, MED test, and Kindred test including 10K, 27K, and 14K videos, respectively. Apart from the Research partition, the two other partitions come with ground truth annotation at video level for 20 event categories, such as *Marriage proposal*, *Attempting bike trick* and *Making sandwich*. In all our experiments, we followed the instruction provided by NIST [16]. We use the MED test and Kindred test partitions to report event

detection results and use the Research partition to train vocabulary concept classifiers.

Video-Concept Annotations: Although the TRECVID Multimedia Event Detection 2013 corpus does not provide video level concept annotations, it comes with a textual summary for each video in the collection, describing what is happening in each video. We use these descriptions to automatically extract video-concept annotations by following the approach proposed in [1]. This approach starts by considering every frequent term in the text collection as a vocabulary concept and terms presence/absence in video descriptions as video-concept annotations. The obtained video-concept annotations are noisy, so [1] proposes to prune the extracted video-concept annotations by excluding some noisy concepts. For this purpose, a classifier is trained and evaluated per concept, then the concepts with low detection accuracy are considered as noise and excluded from the vocabulary. We executed this procedure on the textual descriptions of the Research partition videos, which led to extracting 138 concepts and their video-concept annotations. The extracted concepts cover a wide range of semantics needed to represent events including, objects, actions, scenes and people related concepts. In our experiments, we rely on these extracted annotations as video-concept annotations. Our extracted video-concept annotations are available for download at: *http://www.mediamill.nl/datasets*.

Concept-Event Annotations: Following the standard practice in zero-shot detection, as discussed in Section 3.1, we require concept-event annotations to map the semantic video representation to an event. Similar to video-concept annotations, we rely on textual descriptions of events to extract concept-event annotations. For each event we use 10 textual descriptions from positive examples of the event, provided in the TRECVID Multimedia Event Detection 2013 corpus. Then each description is manually represented in terms of the 138 vocabulary concepts.

4.2 Vocabulary Concept Classifiers

Using the Research partition videos and video-concept annotations we train a classifier per vocabulary concept. As local descriptor we use MBH computed along the motion trajectories [20]. Fisher encoding is used to aggregate them followed by power normalization with $\alpha = 0.2$ as in [6]. This representation is shown to be state-of-the-art for recognizing events using single modality [21]. Better event detection accuracy is obtained by fusing multiple modalities [14] but it is beyond the scope of this paper. We use linear SVM to train the concept classifiers.

4.3 Experiments

1. Zero-Shot Event Detection: We evaluate the effectiveness of our composite concept vocabulary for zero-shot event detection. We compare the discovered OR-composite vocabulary with several baselines: i) primitive concepts vocabulary, which includes 138 concept detectors separately trained per primitive concept. ii) bi-concepts vocabulary, which includes bi-concepts discovered as proposed in [19]. iii) selected concepts vocabulary, which includes a subset of primitive concepts which are more informative per event [11]. Informative concepts are selected using mRMR [17] feature selection from concept-event annotations per event. We select the same number of concepts as in the OR-composite concepts vocabulary.

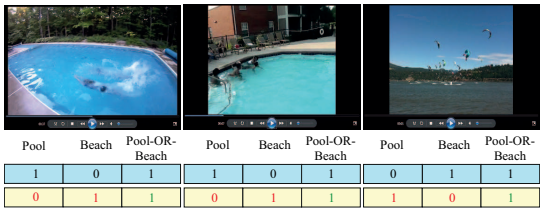

Figure 4: Examples of the predictions made by primitive and composite concept classifiers. Blue and yellow rows denotes the ground-truth and concept classifier predictions, respectively. The primitive concept classifier predictions are incorrect, while the composite concept classifier predictions are correct.

2. OR-Composite vs AND-Composite: We investigate the characteristics of AND-composite and OR-composite concept vocabularies by evaluating their performance for zero-shot event detection. Both the composite vocabularies are discovered as discussed in Section 3.2, by using "*AND*" and "*OR*" operators respectively to compose the concepts. Moreover, we compare their performance with a fused vocabulary containing both the discovered AND-composite and OR-composite concepts.

5. RESULTS

5.1 Zero-Shot Event Detection

The results of this experiment are shown in Table 1. It demonstrates that the OR-composite vocabulary outperforms the primitive vocabulary with mAP of 5.97% vs 4.00% for MED test set and mAP of 14.69% vs 10.47% for Kindred test set. For some events the improvement is considerable, *i.e., winning race without vehicle* and *flash mob gathering*, whose detection accuracy is increased from 2.67% to 10.98% and from 18.41% to 31.86% respectively, on MED test set. By looking into the OR-composite concepts discovered for these events, we find some composite concepts whose classifiers are more accurate than their underlying primitive concept classifiers, as illustrated in Figure 4. Moreover, the quantitative comparison of concept classifier, as reported in Figure 3, demonstrates that OR-composite concepts have higher classification accuracies than their underlying primitive concepts.

We also compare OR-composites vocabulary with bi-concepts vocabulary, which are discovered as proposed in [19]. Our experiments demonstrate that the discovered bi-concepts have poor performance in zero-shot event detection. For some events, *i.e., birthday party* and *flash mob gathering*, bi-concepts vocabulary are even outperformed by the primitive concepts baseline. It is mainly because in [19], bi-concepts are discovered only based on their predictability and without consideration of their effectiveness for detecting events. However, for some concepts, although their composition leads to more predictable bi-concept, the derived bi-concept is less effective in detecting the event. For example in the *birthday party* event, two concepts "dancing" and "indoor" are composed as a bi-concept because detecting dancing in indoor scenes is more accurate. However, this bi-concept is incapable of detecting dancing in outdoor scenes, which leads to missing several birthday party videos which are outdoors. It demon-

Table 1: Experiment 1: Comparing the effectiveness of various concept vocabularies in zero-shot event detection. Our proposed OR-composite concepts outperform present-day alternatives.

	MED test				Kindred test			
Event	Primitive	Selection [11]	Bi-Concepts [19]	OR-Composite	Primitive	Selection [11]	Bi-Concepts [19]	OR-Composite
Birthday party	5.30	4.97	4.72	**7.55**	6.51	4.84	6.49	**9.64**
Changing vehicle tire	0.96	0.97	0.80	**1.81**	**1.21**	1.20	1.33	1.12
Flash mob gathering	18.41	22.98	8.95	**31.86**	15.80	13.67	11.59	**22.52**
Getting vehicle unstuck	3.55	3.40	3.08	**5.54**	1.51	1.54	**2.55**	2.2
Grooming animal	0.91	0.91	0.86	**0.91**	**17.11**	17.06	10.91	17.06
Making sandwich	7.39	7.74	7.39	**7.92**	64.20	62.62	64.20	**66.85**
Parade	19.81	21.90	19.32	**22.36**	5.90	**7.37**	5.96	6.26
Parkour	0.60	0.50	0.91	**2.09**	1.01	0.77	1.00	**4.31**
Repairing appliance	1.06	1.24	0.88	**2.49**	10.81	12.31	5.92	**40.08**
Working sewing project	1.34	1.41	1.36	**1.45**	24.61	29.95	**30.87**	27.32
Attempting bike trick	1.09	1.05	0.62	**2.02**	1.30	1.33	**5.86**	2.78
Cleaning appliance	0.47	0.46	0.47	**0.63**	10.21	7.63	10.15	**23.87**
Dog show	0.10	0.11	**0.25**	0.11	0.61	**0.72**	0.29	0.65
Giving directions location	0.79	0.75	0.52	**2.49**	0.30	0.23	0.23	**0.45**
Marriage proposal	0.13	0.12	**0.23**	0.15	0.91	0.78	0.30	**1.48**
Renovating home	0.55	0.62	0.64	**2.28**	4.20	6.72	4.92	**6.72**
Rock climbing	13.96	14.23	13.94	**14.60**	30.80	30.70	26.58	**34.87**
Town hall meeting	0.52	0.96	0.55	**1.47**	0.71	**1.62**	0.63	0.37
Winning race without vehicle	2.67	3.11	3.09	**10.98**	4.21	5.70	9.05	**15.36**
Working metal crafts project	0.41	0.41	0.49	**0.59**	7.41	7.37	9.64	**9.94**
mean	4.00	4.39	3.45	**5.97**	10.47	10.71	10.42	**14.69**

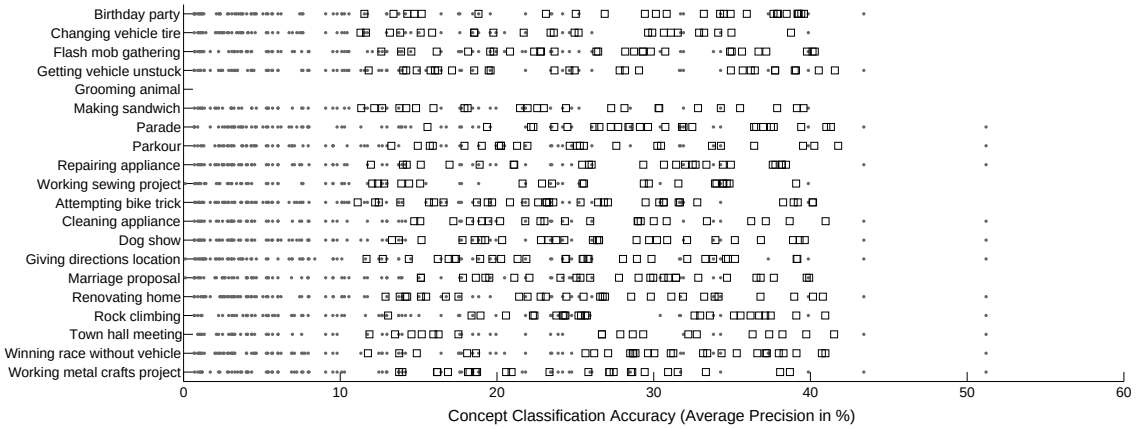

Figure 3: Comparing accuracy of OR-composite concept classifiers and primitive concept classifiers. For each event, the discovered OR-composite concepts are indicated by squares and their underlying primitive concepts are indicated by dots. OR-composite concepts are more accurately classified than their primitive concepts.

strates the importance of considering event predictability in discovering the composite concepts.

Our work is also comparable to concept selection [11], where the vocabulary is obtained by selecting a subset of informative concepts per event. As Table 1 shows, OR-composite vocabulary outperforms the concept selection vocabulary with mAP of 5.97% vs 4.39% for MED test set and 14.69% vs 10.71% for Kindred test set. For some events the improvement is substantial *i.e., winning race without vehicle* event, where OR-composites vocabulary outperforms concept selection by 10.98% versus 2.67% in terms of AP. Looking into the selected concepts for this event, we observe that although the selected concepts are informative for the event, their concept classifiers are inaccurate. More specifically, the accuracy of the classifiers for "running", "jumping", and "walking" concepts, which are among the 10 most informative concept for this event, are 3%, 7%, and 1%, respectively. It demonstrates the importance of considering concept classifiers predictability in discovering the composite concepts.

In summary, we conclude that OR-composite concepts outperform primitive concepts in zero-shot event detection. We explain it based on the observation that OR-composite concepts have more accurate concept classifiers compared to primitive concepts. Moreover, by comparing OR-composite concepts to bi-concepts and selected concepts, we demonstrate the importance of jointly considering concept predictability and event predictability in constructing the concept vocabularies.

5.2 OR-Composite vs AND-Composite

We report the results of this experiment in Table 2. It demonstrates that the OR-composite vocabulary obtains a higher event detection accuracy, 5.97% vs 4.97% for MED test and 11.49% vs 14.69% for Kindred test sets. However, by comparing concept classifier accuracies for AND-composite and OR-composite concepts, as shown in Figure 6, we observe that AND-composite concepts have more accurate concept classifiers, which contradicts with their lower event detection performance. More specifically, the

Table 2: Experiment 2: OR-composite vocabularies outperform AND-composite vocabularies and the best result is obtained by fusing both compositions.

	MED test			Kindred test		
Event	AND-composite	OR-composite	Fusion	AND-composite	OR-composite	Fusion
Birthday party	5.60	7.55	7.55	6.63	9.64	9.64
Changing vehicle tire	1.73	1.81	1.83	1.06	1.12	1.08
Flash mob gathering	26.33	31.86	37.26	17.79	22.52	28.15
Getting vehicle unstuck	3.55	5.54	5.54	1.45	2.20	2.20
Grooming animal	1.02	0.91	0.91	21.89	17.06	17.06
Making sandwich	7.64	7.92	7.92	62.35	66.85	66.85
Parade	23.05	22.36	22.36	5.83	6.26	6.26
Parkour	1.80	2.09	2.18	3.37	4.31	3.92
Repairing appliance	1.52	2.49	2.51	19.80	40.08	48.21
Working sewing project	1.34	1.45	1.45	24.58	27.32	27.32
Attempting bike trick	1.61	2.02	2.16	1.69	2.78	2.39
Cleaning appliance	0.58	0.63	0.80	11.29	23.87	28.54
Dog show	0.13	0.11	0.11	0.22	0.65	0.58
Giving directions location	0.93	2.49	2.49	0.33	0.45	0.45
Marriage proposal	0.13	0.15	0.15	0.55	1.48	1.48
Renovating home	0.60	2.28	2.28	5.83	6.72	6.72
Rock climbing	14.37	14.60	14.68	28.33	34.87	40.72
Town hall meeting	0.52	1.47	1.47	0.73	0.37	0.37
Winning race without vehicle	6.50	10.98	13.59	8.63	15.36	19.25
Working metal crafts project	0.41	0.59	0.59	7.37	9.94	9.94
mean	4.97	5.97	**6.39**	11.49	14.69	**16.06**

concept classification accuracy for AND-composite and OR-composite concepts are 22.45% and 16.19%, respectively in terms of mAP averaged over all concepts and all events. By comparing the number of discovered composite concepts we observe that there are more OR-composite concepts discovered than AND-composite concepts. More specifically, OR-composite vocabulary contains on average 40 OR-composite concepts per event, while AND-composite vocabulary contains on average 12 AND-composite items per event. We explain the higher classification accuracy of AND-composite concepts by the observation that they are more restricted and have less visual diversity in their training examples than OR-composites. For example, the AND-composite "bike-AND-ride" depicts a restricted situation where somebody is riding a bike. But the OR-composite "bike-OR-ride" includes positive examples from various riding actions, *i.e.,* riding bike, horse, skateboard etc., as well as examples from bike in various situations, *i.e.,* riding, repairing, parking etc. In contrast, AND-composite have much less positive examples, which restricts discovering many AND-composite concepts with enough training data to train the classifiers. Moreover, our experiments show that fusing AND-composite and OR-composite concepts in a fused vocabulary, leads to 7.0% and 9.3% relative improvements in MED test and Kindred test sets, respectively. It demonstrates that AND-composite and OR-composite concepts contain complementary information about the events and should both be included in the vocabulary. This hypothesis is validated by looking into the concepts which contribute most to the event detection, as shown in Figure 5. This figure shows that some of the contributing concepts are from OR-composite concepts while others are from AND-composites.

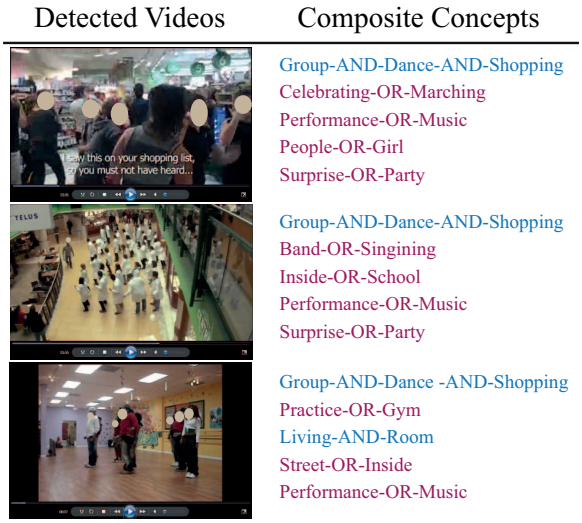

Figure 5: Top three videos detected for *flash mob gathering* event. For each video, we report the five composite concepts which have contributed most to the zero-shot detection of the event. It demonstrates that both the AND-composite and OR-composite concepts are contributing to event detection.

6. CONCLUSIONS

In this paper we propose the notion of composite concepts for zero-shot event detection using concept classifiers. Composite concepts are high order semantics obtained by combining primitive concepts by logical connectors, like "AND" and "OR". We propose an algorithm to automatically discover composite concepts per event that jointly optimizes concept and event predictability. We demonstrate that our discovered composite concepts result in more accurate concept classifiers compared to their underlying primitive concepts, which improves zero-shot event detection accuracy. Moreover, by comparing AND-composite vs OR-composite concepts we observe that AND-composite concepts generally have more accurate concept classifiers. However, AND-composite concepts annotations are sparser than OR-composite concepts, which restricts the number of discovered AND-composite concepts. Moreover, we demonstrate that AND-composite and OR-composite are complementary in representing events and their fusion leads to further improvement.

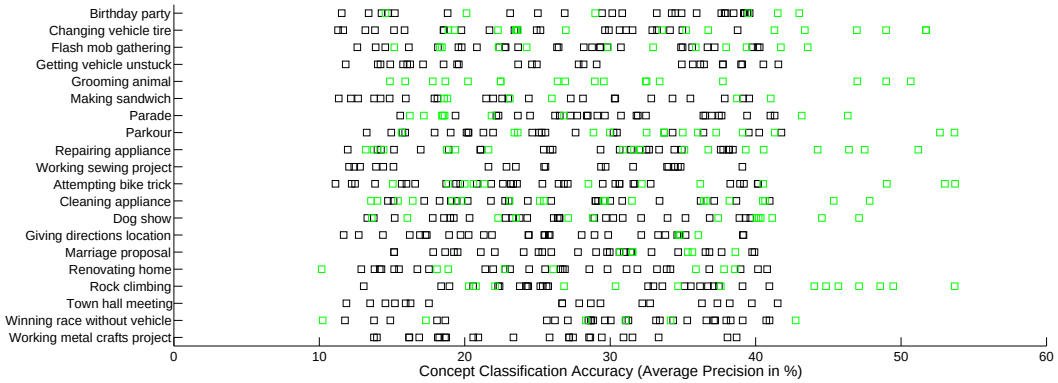

Figure 6: Comparing accuracy of AND-composite and OR-composite concept classifiers. For each event, the discovered AND-composite and OR-composite concepts are indicated by green and black squares, respectively. AND-composite concepts are more accurately classified than OR-composite concepts, but the number of discovered AND-composite concepts are generally less than OR-composite concepts.

Finally, we argue that by training composite concept classifiers, we model the descriptive relationships between concepts explicitly inside the concept classifiers, rather than implicitly in the event classifier. We consider the approach promising for practical use when insufficient event training examples are available to learn concept relationships.

Acknowledgments This research is supported by the STW STORY project and the Dutch national program COMMIT.

7. REFERENCES

[1] T. L. Berg, A. C. Berg, and J. Shih. Automatic attribute discovery and characterization from noisy web data. In *ECCV*, 2010.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM MM*, 2013.

[3] J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. In *CIKM*, 2013.

[4] A. Farhadi and M. A. Sadeghi. Phrasal recognition. *IEEE Trans. PAMI*, 35(12), 2013.

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

[6] M. Jain, H. Jégou, and P. Bouthemy. Better exploiting motion for better action recognition. In *CVPR*, 2013.

[7] L. Jiang, A. G. Hauptmann, and G. Xiang. Leveraging high-level and low-level features for multimedia event detection. In *ACM MM*, 2012.

[8] X. Li, C. Snoek, M. Worring, and A. Smeulders. Harvesting social images for bi-concept search. *IEEE Trans. Multimedia*, 14(4):1091–1104, 2012.

[9] J. Liu, Q. Yu, O. Javed, S. Ali, A. Tamrakar, A. Divakaran, H. Cheng, and H. S. Sawhney. Video event recognition using concept attributes. In *WACV*, 2013.

[10] Z. Ma, Y. Yang, Z. Xu, S. Yan, N. Sebe, and A. G. Hauptmann. Complex event detection via multi-source video attributes. In *CVPR*, 2013.

[11] M. Mazloom, E. Gavves, K. van de Sande, and C. Snoek. Searching informative concept banks for video event detection. In *ICMR*, 2013.

[12] M. Mazloom, A. Habibian, and C. Snoek. Querying for video events by semantic signatures from few examples. In *ICMR*, 2013.

[13] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Trans. Multimedia*, 14(1), 2012.

[14] P. Natarajan, S. Wu, F. Luisier, X. Zhuang, and M. Tickoo. BBN VISER TRECVID 2013 multimedia event detection and multimedia event recounting systems. In *TRECVID Workshop*, 2013.

[15] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *CIVR*, 2006.

[16] P. Over, J. Fiscus, G. Sanders, et al. TRECVID 2012-an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID Workshop*, 2012.

[17] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. PAMI*, 27(8), 2005.

[18] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *CVPR*, 2012.

[19] M. Rastegari, A. Diba, D. Parikh, and A. Farhadi. Multi-attribute queries: To merge or not to merge? In *CVPR*, 2013.

[20] H. Wang, A. Klaser, C. Schmid, and C.-L. Liu. Action recognition by dense trajectories. In *CVPR*, 2011.

[21] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, 2013.

[22] E. Younessian, T. Mitamura, and A. Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In *ICMR*, 2012.

[23] J. Yuan, Z.-J. Zha, Y.-T. Zheng, M. Wang, X. Zhou, and T.-S. Chua. Learning concept bundles for video search with complex queries. In *ACM MM*, 2011.