Semantic Reasoning in Zero Example Video Event Retrieval

M.H.T. de Boer, TNO and Radboud University, The Netherlands
Yi-Jie Lu, City University of Hong Kong, Hong Kong
Hao Zhang, City University of Hong Kong, Hong Kong
Klamer Schutte, TNO
Chong-Wah Ngo, City University of Hong Kong, Hong Kong
Wessel Kraaij, TNO and Leiden University, The Netherlands

Searching in digital video data for high-level events, such as a parade or a car accident, is challenging when the query is textual and lacks visual example images or videos. Current research in deep neural networks is highly beneficial for the retrieval of high-level events using visual examples, but without examples it is still hard to 1) determine which concepts are useful to pre-train (*Vocabulary challenge*); 2) which pre-trained concept detectors are relevant for a certain unseen high-level event (*Concept Selection challenge*). In our paper, we present our Semantic Event Retrieval System which 1) shows the importance of high-level concepts in a vocabulary for the retrieval of complex and generic high-level events and 2) uses a novel concept selection method (*i-w2v*) based on semantic embeddings. Our experiments on the international TRECVID Multimedia Event Detection benchmark show that a diverse vocabulary including high-level concepts improves performance on the retrieval of high-level events in videos and that our novel method outperforms a knowledge-based concept selection method.

CCS Concepts: •**Information systems** → **Query representation; Video search;**

General Terms: Experimentation, Performance

Additional Key Words and Phrases: content-based visual information retrieval, multimedia event detection, zero shot, semantics

1. INTRODUCTION

The domain of content-based video information retrieval has gradually evolved in the previous 20 years. It started as a discipline mostly relying on textual and spoken information in news videos, and moved towards richer multimedia analysis leveraging video, audio and text modalities. The last 10-15 years have shown impressive progress in image classification, yielding larger and larger concept vocabularies. In 2011, the TRECVID MED task defined a testbed for even deeper machine understanding of digital video by creating a challenge to detect high level or complex events, defined as “long-term spatially and temporally dynamic object interactions” [Jiang et al. 2012]. Examples of high-level events are social events (*tailgating party*) and procedural events (*cleaning an appliance*) [Jiang et al. 2012]. Given the extreme difficulty of the MED task, in early years of TRECVID system development was facilitated by providing a set of example videos for the event, making this essentially a supervised video classification task. In the last few years, the MED task has stepped up towards its real challenge: retrieving relevant video clips given -only- a precise textual description of a complex event. In TRECVID MED context, this task is referred to as the zero example case, since no visual examples are provided [Over et al. 2015]. The problem

Author’s addresses: M.H.T. de Boer, TNO, Intelligent Imaging Department, Oude Waalsdorperweg 63, 2597AK, The Hague, The Netherlands and Radboud University, Data Science department, Faculty of Science, Toernooiveld 212, 6525 EC Nijmegen, The Netherlands; maaike.deboer@tno.nl
 1539-9087/YYYY/01-ARTA $15.00
DOI: http://dx.doi.org/10.1145/0000000.0000000

of detecting multimedia events is different from the TRECVID datasets from 2005 to 2008 [Kennedy and Hauptmann 2006; Smeaton et al. 2006]. The TRECVID MED events contain complex and generic high-level events, such as *winning a race without a vehicle*. This query is generic because it is referring to a wide variety of races, including running, swimming, jumping and crawling. The query is also significantly more complex than the entity-based queries, e.g. *emergency vehicle in motion*, used in multimedia research ten years ago, because the number of relevant concepts is higher and the relationship between the concepts plays an important role. Not only should the awareness of a race be captured, but also the winning of a race and the absence of a vehicle in the race (although vehicles could be present on the parking lot near the race or at the side of the street in a marathon).

In our paper, we describe the challenges of building an effective system for zero example complex event retrieval in video. The main issue in zero example video event retrieval is that state of the art machine learning techniques cannot be used, because no training examples are available. A common approach is to use a set of pre-trained classifiers and try to map the event to a set of classifiers. Within this approach two challenges exist: what set of pre-trained classifiers to use (*Vocabulary challenge*) and how to map the event to a set of classifiers (*Concept Selection challenge*).

The Vocabulary Challenge deals with the determination of a good set of concepts to pre-train and put in the vocabulary. This vocabulary is built with pre-trained concept detectors on off-the-shelf datasets. Whereas five to ten years ago fewer than a 1000 pre-trained concepts were available, previous work [Hauptmann et al. 2007b; Aly et al. 2012] was focused on simulations to show how many concepts are actually needed to achieve a reasonable performance. Currently, many datasets with a large vocabulary of pre-trained concepts [Deng et al. 2009; Karpathy et al. 2014; Zhou et al. 2014; Jiang et al. 2015a] are available and we can therefore use actual pre-trained concepts in real datasets instead of simulations. Not all concepts are, however, necessary or useful for a certain test case. For example, the ImageNet dataset [Deng et al. 2009] contains many classes of dog breeds. These concepts are not useful in test cases that only include people and scenes. This implies that it is crucial to at least pre-train and apply those concepts that are valuable for the unseen test case. Some recommendations on how to build a good vocabulary are already available [Habibian et al. 2013].

In this paper, we show the importance of high-level concepts, defined as "complex activities that involve people interacting with other people and/or objects under certain scene" [Chen et al. 2014], because a combination of objects and actions often cannot capture the full semantics of a high-level event. We do not claim that we are the first to use high-level concepts, but we show the difference in performance for different types of concepts.

The Concept Selection challenge embeds the problem that the system has no prior knowledge about the events, so in many cases no precise visual concept detectors are available. Commonly, this challenge is approached by mapping the event to a set of classifiers by optimizing the match between the User Query (*UQ*) and the System Query (*SQ*). Within the TRECVID community, this is also referred to as Semantic Query Generation [Over et al. 2015]. Here the User Query is a textual description of the event and the System Query is a combination of concepts present in our vocabulary. In this paper, we will refer to the term *concept* as the label or name of the concept itself and to *concept detectors* as pre-trained classifiers. In this challenge, we build upon the existing word2vec models [Mikolov et al. 2013; Pennington et al. 2014] which use semantic embeddings. The main novelty of our method is that it accurately selects the proper concepts without the problem of query drift, in which the selected concepts create a drift towards one facet of the query [Carpineto and Romano 2012].

The main contributions of this paper can be summarized as follows:

— We show the importance of high-level concepts in defining a good vocabulary of pre-trained concept classifiers in the case of search queries that contain high-level events.
— We introduce an incremental word2vec method (*i-w2v*) for concept selection that is more robust to query drift and cut-off parameter tuning.

The next section contains related work. We focus on our two challenges. The third section explains our Semantic Event Retrieval System that includes our novelties in both challenges. The fourth section presents the experiments conducted on the international benchmark TRECVID Multimedia Event Detection [Over et al. 2015] and the results are included in Section 5. The sixth section contains a discussion and the final section provides the conclusion.

2. RELATED WORK

In this section we only focus on the Vocabulary challenge and the Concept Selection challenge in zero example video event retrieval.

2.1. Vocabulary

Concept vocabularies are designed as a representation layer for a specific purpose, such as indexing descriptors for video clips, shots or frames. Ideally, concept vocabularies consist of unambiguous precise descriptors of entities, activities, scenes, objects and ideas. Different vocabularies are developed for different purposes. Combining different vocabularies often results in vagueness and ambiguity, such as polysemy and homonymy. We will focus on two properties of concepts: *level of complexity* and *level of granularity*. In the level of complexity, three levels can be differentiated. First, *low-level* concepts are the basic components in images or videos, such as objects. Second, *mid-level* concepts are basic actions, activities or interactions. Actions or activities are a “sequence of movements” [Chen et al. 2014] and can be performed by one entity, such as people or objects. Interactions are actions between two or more entities. Third, *high-level* concepts are “complex activities that involve people interacting with other people and/or objects under certain scene” [Chen et al. 2014]. The key difference between mid-level and high-level concepts is that a high-level concept contains multiple actions and interactions evolving over time [Chen et al. 2014], such as the difference between the action *horse riding* and the event *horse riding competition*. Furthermore, concepts can have different levels of granularity, also referred to as specificity. Examples are animal (*general*), dog and chihuhua (*specific*).

The importance of the level of granularity in a vocabulary was already indicated by Hauptmann et al. [2007b] and Habibian et al. [2013]. Both argue that in video event recognition a mixture of both general and specific concepts achieves higher performance compared to using only general or specific concepts. Interestingly, both papers state that the general concepts achieve in general higher performance compared to the specific concepts, because specific concepts only occur in a few videos, and many general concepts can be distinctive enough to recognize an event. The importance of the level of complexity is not yet introduced, but Habibian et al. [2013] recommend to use a vocabulary that contains concepts of the following categories: object, action, scene, people, animal and attribute. Using our definitions an action is comparable to a mid-level concept and the concepts from the other categories are low-level concepts. Another work of these authors introduces primary concepts and bi-concepts [Habibian et al. 2014a].

Other recommendations from Habibian et al. [2013] are 1) use a vocabulary with at least 200 concepts; and 2) do not use too many concepts of one type, such as animals or people. Additionally, they argue that it is better to include more concepts than to improve the quality of the individual concepts, which is also concluded by Jiang et al.

[2015b]. Previous research of Aly et al. [2012] indicated that few concepts (100) with a simulated detector performance of only 60% is already sufficient to achieve reasonable Mean Average Precision performance (20%). Hauptmann et al. [2007a] argue that 3000 concepts are needed for a Mean Average Precision of 65%. We follow this recommendation and focus on extending the vocabulary instead of improving performance of concept detectors.

In addition to the type of concepts, Jiang et al. [2015b] report the influence of training with different datasets on performance for the events in the TRECVID Multimedia Event Detection task. The dataset with the highest performance is Sports [Karpathy et al. 2014], followed in descending order by the 1000 concepts from ImageNet [Deng et al. 2009], the Internet Archive Created Commons (IACC) dataset [Over et al. 2014], the big Yahoo Flickr Creative Common dataset (YFCC) [Thomee et al. 2015] and the Do It Yourself (DIY) dataset [Yu et al. 2014]. We use the concepts of their top two performing datasets in our vocabulary. Furthermore, one of their recommendations is to train concept detectors on large datasets, both in terms of training examples as well as number of concepts. We take this recommendation into account and focus on large datasets.

2.2. Concept Selection

Many different techniques are used in Concept Selection. Liu et al. [2007] present five categories in which concepts can be selected, of which we use three as a guideline to give an overview of the different methods used in the recent years. The first category is making use of an ontology. These ontologies or knowledge bases can be created by expert (*expert knowledge base*) or created by the public (*common knowledge base*). Expert knowledge bases provide good performance, but dedicated expert effort is needed in the creation of such a knowledge base. Some early work on expert knowledge bases and reasoning in the field of event recognition is explained in Ballan et al. [2011]. One current expert ontology for events is EventNet [Ye et al. 2015]. Common knowledge bases, such as Wikipedia [Milne and Witten 2013] and WordNet [Miller 1995], are freely available and often used in the video event retrieval community [Neo et al. 2006; Yan et al. 2015; Tzelepis et al. 2015], but might not contain the specific information that is needed. A comparison of performance between an expert knowledge base and two common knowledge bases, which are Wikipedia and ConceptNet, is given in de Boer et al. [2015]. Concept selection in common knowledge bases is often done by using the most similar or related concepts to events found in the knowledge base. An overview of the type of methods to find similar or related concepts can be found in Natsev et al. [2007]. The number of selected concepts and the similarity measures used differ per paper and no conclusive result on which method works best is found.

The second category is making use of machine learning techniques. Machine learning techniques can be used to automatically select the proper concepts. These techniques are used more often in tasks with example videos, because many models need training examples. In the zero example video event retrieval, graphical models such as hidden Markov models [Dalton et al. 2013], are used. More often statistical methods are used, such as co-occurrence statistics [Mensink et al. 2014] and a skip-gram model [Chang et al. 2015]. One group of current state of the art models is word2vec, which produce semantic embeddings. These models either use skip-grams or continuous bag of words (CBOW) to create neural word embeddings using a shallow neural network that is trained on a huge dataset, such as Wikipedia, Gigawords, Google News or Twitter. Each word vector is trained to maximize the log probability of neighboring words, resulting in a good performance in associations, such as *king - man + woman = queen*. Two often used models are the skip-gram model with negative sampling (SGNS) [Mikolov et al. 2013], which has relations to the pointwise mutual information [Levy and Goldberg 2014], and the Glove model [Pennington et al. 2014], which uses a factorization of the log-count matrix. Although Pennington et al. [2014] claimed to have performance superior to SGNS, this is highly debated by Levy et al. [2015] and Goldberg[1]. The advantage of word2vec over other semantic embedding methods is that the latent variables are transparent, because the words are represented in vector space with only a few hundred dimensions. Examples of other semantic embedding methods are Wu et al. [2014] with their common lexicon layer and Habibian et al. [2014b] with VideoStory and Jain et al. [2015] with the embedding of text, actions and objects to classify actions.

The third category is making use of relevance feedback. User clicks or explicit relevance judgements from users can be used to optimize the results. A review of relevance feedback in content based image retrieval can be found in Patil and Kokare [2011]. In concept selection using relevance feedback often an initial set of concepts is chosen using the ontology, machine learning techniques or one of the other techniques and a user is asked to remove the irrelevant concepts and/or to adjust the importance of concepts [Jiang et al. 2015b; Chang et al. 2015]. A second option is to refine the text query instead of removing concepts [Xu et al. 2015]. A third option is to use weakly labelled data [Chang et al. 2016] to dynamically change the weights of the selected concepts. Besides user interaction, pseudo-relevance feedback can be used. In pseudo-relevance feedback we assume that the top videos are relevant for the query [Jiang et al. 2014a; Jiang et al. 2014b]. Although this method by the CMU team has top performance in TRECVID MED 2014, pseudo-relevance feedback is a high risk for rare events. In our experiments, we focus on the first run of the video event retrieval system and, therefore, do not include pseudo-relevance feedback. We, however, compare our method with a method that uses a user to create the System Query.

In addition to the different categories from Liu et al. [2007] and Jiang et al. [2015b] found that a sensible strategy for concept selection might be to incorporate more relevant concepts with a reasonable quality. They state that automatic query generation or concept selection is still very challenging and combining different mapping algorithms and applying manual examination might be the best strategy so far. Huurnink et al. [2008] propose a method to asses the automatic concept selection methods and compare that method to a human assessment. Mazloom et al. [2013] show that an informative subset of the vocabulary can achieve higher performance than just using all concepts of the vocabulary in a setting of video event retrieval with examples. This strategy is also used in our previous work [Lu et al. 2016] that uses evidential pooling of the concepts in the video.

3. SEMANTIC EVENT RETRIEVAL SYSTEM

In our Semantic Event Retrieval System we use five large external datasets to form our vocabulary, which is explained in the following subsection. Our vocabulary is used in our concept selection method to transform the user query (UQ) into a System Query (SQ), as explained in the second subsection. UQ is a fixed textual description of an event, for which we only use the name of the event. SQ is a list of concepts (c) and their associated similarities (c_s). The constraints on our SQ are: *sparsity*, *non-negativity* and *linear weighted sum*. Regarding sparsity, we use an informative subset of concepts, as recommended by Mazloom et al. [2013] and similar to our previous findings, resulting in a sparse set of concepts in SQ. No negative similarities are used, because in our findings this decreases performance. For example, in the event *winning a race without a vehicle* using a negative similarity for the concept *vehicle* decreases performance,

[1]On the importance of comparing apples to apples: a case study using the GloVe model, Yoav Goldberg, 10 August 2014

because in some videos of this event a parking lot with vehicles is present at the beginning of the video. The linear weighted sum is used to combine the concepts in our SQ to create the event score for a certain video ($S_{e,v}$). The concept detector score per video ($c_{d,v}$) is the concept detector score (d) belonging to a video (v).

The formula to create the event score is shown in Equation 1.

$$S_{e,v} = \sum_{c \in SQ} c_s \cdot c_{d,v}, \tag{1}$$

, where c is the concept, V is the vocabulary, c_s is the similarity of concept c, $c_{d,v}$ is the concept detector score for concept c over video v. The event scores can be used to order the videos and calculate performance.

3.1. Vocabulary

While creating the vocabulary, we follow the recommendations of Habibian et al. [2013], which are to use a large and diverse vocabulary, and use the top two performing datasets from Jiang et al. [2015b], i.e. Sport and ImageNet. Furthermore, we aim for a set of datasets that not only contains low- and mid-level concepts, but also high-level concepts. Figure 1 shows our interpretation of the different datasets on the level of complexity.

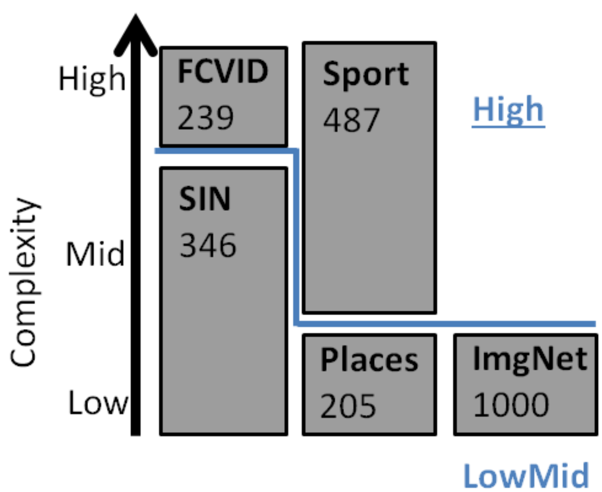

Fig. 1. The level of complexity for the five datasets used in this paper.
The number under each dataset indicates the amount of concepts in the dataset.

The two low-level datasets are ImgNet [Deng et al. 2009] and Places [Zhou et al. 2014]. ImgNet, which is an abbreviation for ImageNet, contains low-level objects and for our vocabulary the standard subset of 1000 objects is used. The Places dataset does not contain objects, but scenes or places. We have one dataset that contains both low- and mid-level concepts: SIN [Over et al. 2015]. These concepts have been developed for the TRECVID Semantic Indexing Task of 2015. We also included one dataset that contains both mid-level and high-level concepts: Sport [Karpathy et al. 2014]. This is a dataset that contains one million sports videos, classified into 487 categories. Our high-level dataset is the Fudan Columbia Video dataset [Jiang et al. 2015a], which contains 239 classes within eleven high-level groups, such as art and cooking&health.

Table I. Overview Datasets

Name	#Concepts	Structure	Dataset
FCVID	239	DCNN+SVM	Fudan-Columbia [Jiang et al. 2015a]
SIN	346	DCNN	TRECVID SIN [Over et al. 2015]
Sport	487	3D-CNN [Tran et al. 2014]	Sports-1M [Karpathy et al. 2014]
Places	205	DCNN	MIT Places [Zhou et al. 2014]
ImgNet	1000	DCNN [Krizhevsky et al. 2012]	ImageNet [Deng et al. 2009]

Table I shows additional information on the datasets, such as the amount of concepts, the reference to the publication of the dataset and the structure used to train the concept detectors. Training of the concepts is done by using one of the states of the art DCNN architectures. The original DCNN architecture of Krizhevsky et al. [2012], named AlexNet, is used for ImgNet. The output of the eighth layer of the DCNN network trained on the ILSVRC-2012 [Deng et al. 2009] is used as concept detector score per keyframe. This DCNN architecture is fine-tuned for both SIN and Places. The concept detector scores per keyframe are max pooled to obtain the score per video. The keyframes are extracted at the rate of one keyframe per two seconds.

The two high-level datasets are annotated on video level instead of keyframe level and are, therefore, trained in a slightly different way. FCVID also uses the same DCNN architecture, but the seventh layer of the network is used as an input for an SVM. This SVM is trained on the videos within the dataset on video level instead of keyframe level. The Sport dataset is trained with the 3D CNN network of Tran et al. [2014].

3.2. Concept Selection (i-w2v)

Our incremental word2vec method (*i-w2v*) starts with a vector containing the words in the User Query (UQ). In our experiments, the UQ is the name of an event, such as [parking, vehicle]. On the other hand, we have a vocabulary with concepts. These concepts can also be represented as a vector, such as the concept [police, car]. In the function *sim(c,UQ)*, we use the Gensim code[2], which is an implementation of the SGNS model [Mikolov et al. 2013], to calculate the cosine similarity between UQ and each of the concepts in the vocabulary. This similarity is stored in c_s. We sort the concepts in the vocabulary based on this similarity. We discard the concepts with a similarity less than 80% of the highest similarity. This cut-off is used to decrease the possibility of introducing noise. Subsequently, we try whether a combination of concepts will increase the similarity to take care of the query drift. Where other methods might only choose the top five as the selected concepts, we - only - include the concepts that increase the similarity. In the multidimensional word2vec space, one facet might have a vector into one direction towards UQ, whereas another facet might have a vector into another direction. Using both concepts will move the vector more towards the vector of UQ and increase the cosine similarity. We start with using the concept with the highest similarity in a concept vector. We iteratively add concepts (in order of their similarity) to this concept vector and each time compare the cosine similarity of the new vector to UQ. If the similarity is higher with the concept than without, we retain the concept in the concept vector. In the case of the event *parking a vehicle*, the first concept is *vehicle*. All types of vehicle, such as *police car* or *crane vehicle* are not added to the concept list as the concept list with the police car added, such as [vehicle, police, car] does not increase the cosine similarity to UQ. The concept *parking lot*, which was not in the top five concepts, is included, because the facet *vehicle* and the facet *parking (lot)* together increase the similarity to the event *parking a vehicle*. Similarly, the tenth concept *parking meter* is not included as it covers the same facet as *parking lot*.

[2]https://radimrehurek.com/gensim/models/word2vec.html

The output of the Concept Selection method is the list of selected concepts and their original cosine similarity c_s to UQ. This concept selection method has a complexity of $O(n)$ in which n is the amount of concepts, because we have to calculate the similarity between the query and each of the concepts. This method is faster than look-up time of the video in the database, which makes it applicable for real-time systems.

Table II shows that our method is robust to a range of cut-offs, both percentages and a fixed similarity threshold of 0.1, on the vocabulary using pre-trained concepts from all datasets mentioned in the previous section (referred to as the All vocabulary). The average amount of concepts remaining after applying our algorithm is also included in Table II. The novelty in our method is to only add the concepts that improve the similarity to the full event. To our knowledge, current word2vec models did not yet look into solutions to a possible query drift in this way.

Table II. MAP performance for different cut-off points in i-w2v algorithm (All vocabulary on MED14Test)
Cut-off means discard all concepts that have a similarity lower than the cut-off value compared to the concept with the highest similarity

Cut-off	MAP	Average Number of Concepts
none	0.136	9.4 ± 13.4
25%	0.136	9.3 ± 13.4
50%	0.137	7.2 ± 12.1
75%	0.141	3.8 ± 6.3
80%	0.142	3.0 ± 4.6
85%	0.142	2.3 ± 2.4
90%	0.142	1.9 ± 1.3
0.1	0.142	2.9 ± 5.3

4. EXPERIMENTS

In our experiments, we use the MED2014Test Set of the TRECVID Multimedia Event Detection Pre-specified Zero-Example task of 2015 [Over et al. 2015]. The MED2014Test contains more than 27,000 videos and has ground truth information for twenty events. The evaluation metric is Mean Average Precision [Over et al. 2015]. All video scores are sorted in descending order and the rank of the positive videos is used in the evaluation. The next sections explain our experiments on the Vocabulary Challenge and Concept Selection challenge.

4.1. Vocabulary

In the experiments on the Vocabulary challenge, we compare performance of vocabularies that consist of 1) only one dataset; 2) only low- and mid-level concepts (*LowMid*); 3) only high-level concepts (*High*); 4) low-, mid- and high-level concepts (*All*). The datasets used in the LowMid, High and All vocabularies are visualized in Figure 1 on the previous page.

According to the literature, combining resources generally improves robustness and performance and therefore we hypothesize that 1) *All* outperforms all other vocabularies. Our intuition is that the high-level concepts play an important role in the detection of high-level events and thus we hypothesize that 2) *High* outperforms *LowMid* and 3) Sport and FCVID outperform the other single datasets.

The Concept Selection method used for the experiments on the Vocabulary Challenge is not our proposed Concept Selection method, but the best number of concepts over all events (top-k) using the original word2vec method. This number is determined by experiments on the MED2014 TEST with a varying number of selected concepts, from one to twenty. This number therefore displays the best possible k over all events for these twenty events and is thus not influenced by the proposed Concept Selection method, enabling an independent experiment on the vocabularies.

4.2. Concept Selection

In the experiments on the Concept Selection challenge, we compare performance of our proposed Concept Selection method (*i-w2v*) to the original word2vec method (*top-k*), a knowledge-based method (*CN*), a method using manually selected concepts and weights (*manual*) and the currently known state of the art methods describing their performance on MED14Test. Relating back to the related work, CN is selected as a method from the first category (*ontology*). The i-w2v method falls within the second category (*machine learning*), and the manual method falls within the third category (*relevance feedback*). We hypothesize that 1) i-w2v outperforms CN and 2) manual outperforms both CN and i-w2v. This second hypothesis is based on the finding of Jiang et al. [2015b] that automatic Concept Selection is still a challenge.

In the *CN* method, UQ (event name) is first compared to the concepts in the vocabulary. If a concept completely matches UQ, this concept is put in SQ. If no concept completely matches UQ, ConceptNet is used to expand UQ. In this expansion, ConceptNet 5.3 is automatically accessed through the REST API and all words with the relation *RelatedTo*, *IsA*, *partOf*, *MemberOf*, *HasA*, *UsedFor*, *CapableOf*, *AtLocation*, *Causes*, *HasSubEvent*, *CreatedBy*, *Synonym* or *DefinedAs* to UQ are selected, split into words by removing the underscore and compared to the lemmatized set of concepts in the vocabulary. The matching concepts are put in the SQ. The value for c_w is determined by the following equation:

$$c_w = \left(\frac{score_{rel}}{30}\right)^3 \tag{2}$$

This equation is based on the experiments in de Boer et al. [2015], where they explain that the scores are often between zero and thirty, which would create a value between zero and one. The third power is based on previous experiments and has some ground in Spagnola and Lagoze [2011], because they explain that ConceptNet uses the third root of the score of the edges to calculate the final score.

If the query expansion directly to UQ still gives no related concepts, the separate words in UQ are compared to the concepts. The words with a matching concept are put in SQ and the other words are expanded through ConceptNet. In order to avoid query drift, the sum of the weights of the expanded words should be the same as the weight of a matched concept. If for example UQ contains of two words, each set of concepts that represent one word should have a weight of 0.5.

In the *manual* method a human researcher had to select the relevant concepts and weights for those concepts for each event. The researcher was presented the event description provided within the TRECVID MED [Over et al. 2015] benchmark, access to the internet to search for examples for the event and knowledge sources such as Wikipedia or the dictionary and the list of concepts. In order to help the human researcher, the ranked list from our i-w2v method (without similarities) was provided to show a list that is somewhat ordered in terms of relevance to the event. This human researcher is a non-native fluent English speaker with a West-European background. The human researcher was instructed to create a diverse and concise list of concepts, to prevent query drift and adding too much noise. The human researcher had to provide weights for the concepts that summed up to one.

5. RESULTS

5.1. Vocabulary

The results of the Average Precision performance of the different vocabularies are shown in Table III. The bold number indicates the highest performance per event per vocabulary, both from the vocabularies that contain a single dataset and the vocabularies with concepts from multiple datasets.

Comparing performance of All to the other datasets, we clearly see that on average the combination of all resources is better than using a subselection of the resources, which is consistent with our first hypothesis. Additionally, LowMid and High both have a performance which is on average higher than any of the single dataset vocabularies in that category.

Table III. Average Precision per Vocabulary using top-k word2vec concept selection (k is optimal determined on MED2014TEST).
Bold is highest in row and group.

	ImgNet (1)	Places (1)	SIN (1)	Sport (1)	FCVID (1)	LowMid (2)	High (1)	All (1)
AttemptBikeTrick	0.061	0.002	0.050	0.003	**0.062**	**0.078**	0.062	0.062
CleanAppliance	0.011	0.011	0.009	0.006	**0.062**	0.009	**0.062**	**0.062**
DogShow	0.013	0.011	0.011	**0.766**	0.006	0.005	**0.766**	**0.766**
GiveDirection	**0.006**	0.001	0.002	0.002	0.001	**0.006**	0.002	**0.006**
MarriageProposal	0.002	0.002	0.003	0.002	**0.010**	0.003	**0.010**	**0.010**
RenovateHome	**0.003**	0.002	0.002	0.002	0.001	**0.002**	0.001	0.002
RockClimbing	**0.003**	0.004	0.005	**0.128**	0.065	0.003	**0.128**	**0.128**
TownHallMeeting	0.001	0.008	0.015	0.001	**0.148**	0.015	**0.148**	**0.148**
WinRace	0.006	0.005	0.006	0.010	**0.011**	0.005	**0.010**	0.005
WorkMetalCraftsProject	0.003	0.003	0.002	0.001	**0.005**	0.003	**0.005**	**0.005**
Beekeeping	**0.620**	0.013	0.007	0.011	0.262	**0.620**	0.262	**0.620**
WeddingShower	0.002	0.002	0.002	0.003	**0.005**	0.002	**0.005**	0.002
VehicleRepair	0.002	0.003	0.006	**0.007**	0.001	0.002	0.001	**0.006**
FixMusicalInstrument	0.021	0.024	0.001	0.002	**0.147**	0.004	**0.147**	**0.147**
HorseRidingCompetition	0.022	**0.224**	0.071	0.044	0.098	**0.224**	0.098	0.044
FellingTree	0.002	0.052	0.019	0.012	**0.026**	0.002	**0.026**	**0.026**
ParkingVehicle	0.003	0.023	0.022	0.002	**0.217**	0.023	**0.217**	**0.217**
PlayingFetch	0.001	0.001	**0.001**	0.001	0.001	**0.001**	0.001	0.001
Tailgating	0.004	0.010	0.002	0.002	**0.232**	0.006	**0.232**	**0.232**
TuneMusicalInstrument	0.035	**0.052**	0.004	0.001	0.052	0.008	**0.052**	**0.052**
MAP	0.041	0.023	0.012	0.050	**0.071**	0.051	0.112	**0.127**

Furthermore, the high-level concepts are important in these experiments, because High outperforms LowMid and the high-level datasets Sports and FCVID outperform Places and SIN. Besides the complexity of the datasets, the amount of concepts could also be a factor. A higher amount of concepts increases the possibility that the event can be captured within these concepts. This factor can be further verified by the plot in Figure 2.

In this plot, the correlation between the amount of concepts for each of the complexities is shown. LowMid has a high correlation, whereas High has not (R^2 LowMid = 0.867 and R^2 High = 0.412) between amount of concepts and MAP. The plot clearly shows that High performs better than LowMid with the same amount of concepts.

Please note that these results could also be explained by that the high level concepts are trained in a domain more like TRECVID MED compared to the domain in which the low level concepts are trained. This domain shift could decrease the performance of the low level concepts compared to the high level concepts.

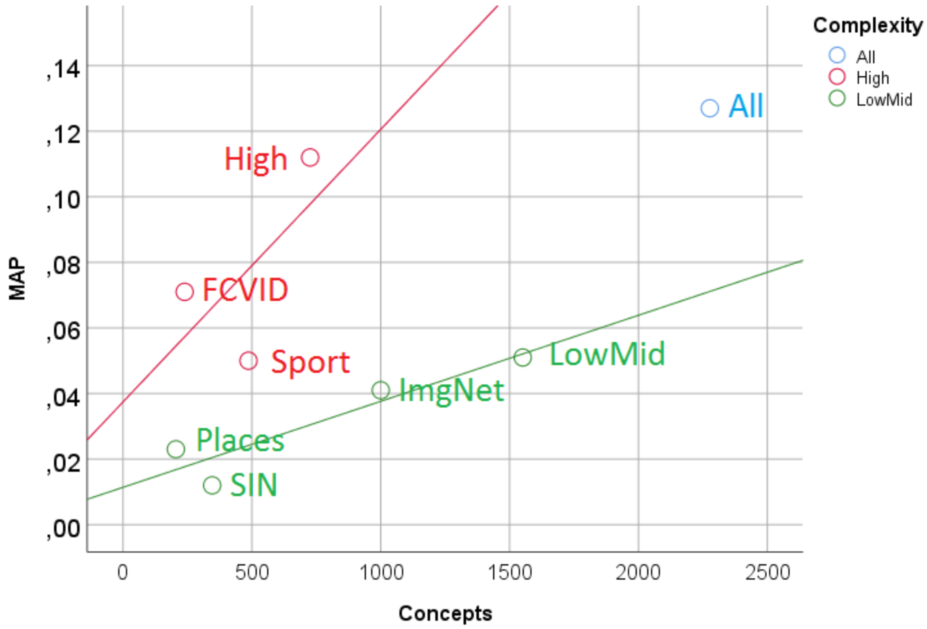

Fig. 2. Correlation between Amount Concepts and MAP for different complexities

5.2. Concept Selection

The previous section shows the top-k performance for different vocabularies, whereas in this section we compare the Concept Selection methods. The Average Precision performance results for our Concept Selection experiments are shown in Table IV. The bold number indicates the highest performance per event per vocabulary. The italic numbers for the CN method indicate random performance, because no concepts are selected. In the All vocabulary, for some events performance of all concept selection methods is equal, indicating that a complete match between the event and a concept in the vocabulary is found. In each of the methods a complete match will result in only selecting that concept. These events are, therefore, displayed on top of the table and separated from the 'interesting' events on the bottom of the table.

Additionally, we compare our best performance against state of the art performance reported on the same dataset in Table V. Performance of CN, top-k and i-w2v on the All vocabulary is shown. This performance is directly comparable to EventPool, because the same vocabularies are used. The vocabularies used by Chang et al. [2016] and Jiang et al. [2015b] are comparable in size and type of concepts. In Bor, PCF and DCC semantic concepts are discovered using weakly labelling the TRECVID MED research set using word2vec vectors. Bor uses borda rank to aggregate the weights on the concepts. PFC uses a pair-comparison framework. DCC uses a dynamic composition to determine the appropriate weights. Fu is the AND-OR method proposed by Habibian et al. [2014a] to create an AND-OR graph of the concepts, but applied to the vocabulary of Chang et al. [2016]. The vocabulary of Habibian et al. [2014a] was composed of 138 concepts. These concepts were automatically extracted from the TRECVID MED research set. Jiang et al. [2015b] uses an average fusion of the mapping algorithms that use exact word matching, Wordnet, Pointwise Mutual Information and word embeddings. Table V shows a gain in MAP of 1% compared to state of the art methods.

Table IV. Average Precision on MED2014TEST for proposed i-w2v, top-k, ontology-based CN and manual concept selection Top part are events with direct matches to a concept. Bold is highest value in row and group.

	LowMid				**High**				**All**			
	i-w2v	top-k	CN	manual	i-w2v	top-k	CN	manual	i-w2v	top-k	CN	manual
AttemptBikeTrick	**0.103**	0.078	0.021	0.08	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062
CleanAppliance	0.014	0.009	0.005	**0.021**	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062
DogShow	**0.021**	0.005	0.011	0.011	0.766	0.766	0.766	0.766	0.766	0.766	0.766	0.766
MarriageProposal	0.003	0.003	*0.001*	**0.005**	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
RockClimbing	0.006	0.003	0.002	**0.025**	0.309	0.128	0.309	0.309	0.309	0.128	0.309	0.309
TownHallMeeting	0.012	0.015	0.007	**0.023**	0.148	0.148	0.148	0.148	0.148	0.148	0.148	0.148
FixMusicalInstrument	0.025	0.004	0.009	**0.057**	0.147	0.147	0.147	0.147	0.147	0.147	0.147	0.147
Tailgating	0.007	0.006	0.002	**0.010**	0.232	0.232	0.232	0.232	0.232	0.232	0.232	0.232
MAP (direct)	0.024	0.015	0.008	**0.029**	0.217	0.194	0.217	0.217	0.217	0.194	0.217	0.217
GiveDirection	**0.006**	**0.006**	0.002	0.004	0.002	0.002	0.001	**0.004**	0.006	0.006	0.002	**0.008**
RenovateHome	0.003	0.002	**0.017**	0.003	0.001	0.001	0.002	**0.015**	0.002	0.003	**0.015**	0.008
WinRace	0.006	0.005	0.007	**0.035**	0.043	0.01	0.007	**0.093**	0.021	0.005	0.011	**0.093**
WorkMetalCraftsProject	0.003	0.003	0.001	**0.016**	0.005	0.005	0.001	**0.007**	0.005	0.005	0.001	**0.008**
Beekeeping	0.62	0.62	0.65	**0.694**	**0.262**	**0.262**	**0.262**	**0.262**	0.62	0.62	0.666	**0.714**
WeddingShower	0.002	0.002	0.002	**0.005**	**0.005**	**0.005**	0.002	**0.005**	0.005	0.002	0.002	**0.005**
VehicleRepair	**0.006**	0.002	0.003	**0.006**	0.006	0.001	0.003	**0.162**	0.006	0.006	0.005	**0.284**
HorseRidingCompetition	0.182	**0.224**	0.015	0.183	0.098	0.098	0.096	**0.261**	0.119	0.044	0.096	**0.288**
FellingTree	**0.031**	0.002	0.006	0.015	0.024	0.026	0.001	**0.033**	**0.042**	0.026	0.008	0.015
ParkingVehicle	0.026	0.023	0.022	**0.031**	**0.217**	**0.217**	0.001	**0.217**	**0.220**	0.217	0.013	0.216
PlayingFetch	0.001	0.001	**0.012**	0.004	0.001	0.001	0.022	**0.023**	0.001	0.001	0.02	**0.023**
TuneMusicalInstrument	**0.050**	0.008	0.012	0.046	**0.052**	**0.052**	*0.001*	**0.052**	**0.052**	**0.052**	0.012	**0.052**
MAP (no direct matches)	0.078	0.075	0.062	**0.087**	0.06	0.057	0.033	**0.095**	0.092	0.082	0.071	**0.143**
MAP (all)	0.056	0.051	0.04	**0.064**	0.123	0.112	0.107	**0.144**	0.142	0.127	0.129	**0.173**

Table V. Comparison to State of the Art (MAP reported on MED2014TEST)

Method	**MAP**
AND-OR[Habibian et al. 2014a]	0.064
Bor[Chang et al. 2016]	0.102
Fu [Chang et al. 2016; Habibian et al. 2014a]	0.111
PCF [Chang et al. 2016]	0.114
AutoSQGSys [Jiang et al. 2015b]	0.115
top-k (All)	0.127
EventPool [Lu et al. 2016]	0.129
CN (All)	0.129
DCC [Chang et al. 2016]	0.134
i-w2v (All)	**0.142**

Comparing the Concept Selection methods, manual is the best overall Concept Selection method, as expected by our hypothesis. The largest differences between manual and i-w2v and CN are in *VehicleRepair* and *HorseRidingCompetition* in High and All. Table VI shows the different concepts and similarities for *VehicleRepair* in All. The concept *assemble bike* has high performance, because this is the only concept that differs between i-w2v / top-k and manual. In the High vocabulary, performance for this event drops, because the concept *vehicle* is no longer within the vocabulary. This same phenomenon happens in the event *Beekeeping* with the concept *apairy*. The main difference in performance in *HorseRidingCompetition* is that the human researcher was able to select all types of horse riding competitions, whereas CN only selected *dressage* and i-w2v only selected the concept *horse racing* in High and *horse racing* and *horse* in All. The difference between High and All with manual in this event is due to the concept *horse race course*.

Table VI. Comparison for VehicleRepair in All

i-w2v / top-k		**CN**		**manual**	
c	c_s	c	c_s	c	c_s
vehicle	0.760	vehicle	0.500	vehicle	0.5
		band aid	0.095	assemble bike	0.5
		highway	0.095		
		apartments	0.095		
		boating	0.095		
		shop	0.095		
		casting fishing	0.024		

Following our hypothesis, i-w2v outperforms CN in all vocabularies. I-w2v even outperforms manual in some events, of which *FellingTree* is the most interesting. Table VII shows the concepts and similarities of the different methods for the event *FellingTree* in All. In i-w2v, the concept *tree farm* provides for high performance, whereas *chain saw* decreases performance compared to only using the concept *fruit tree pruning*. In CN, the wrong expansion from *felling* to *falling* to all concepts, except for *trees*, causes the low performance. Please note that the human researcher has the highest performance in High. The selected concepts for manual in High are *forest* and *fruit tree pruning*.

Table VII. Comparison for FellingTree in All

i-w2v		CN		manual		top-k	
c	c_s	c	c_s	c	c_s	c	c_s
fruit tree pruning	0.720	trees	0.500	trees	0.5	fruit tree pruning	0.720
tree frog	0.686	cliff	0.186	chain saw	0.5		
tree farm	0.678	painting	0.106				
		skateboarding	0.085				
		climbing	0.040				
		windows	0.040				
		head	0.002				
		running	0.001				
		building	7×10^{-6}				

Table VIII. Comparison for RenovateHome in LowMid

i-w2v		CN		manual		top-k	
c	c_s	c	c_s	c	c_s	c	c_s
apartment building-outdoor	0.542	apartments	0.113	apartment building-outdoor	0.25	apartment building - outdoor	0.542
building	0.526	city	0.102	apartments	0.25		
home office	0.475	person	0.083	construction site	0.5		
apartments	0.466	wardrobe	0.065				
church building	0.465	sofa	0.065				
building facade	0.452	tabby cat	0.065				
mobile home	0.437	closet	0.065				
		bedroom	0.065				
		comfort	0.065				
		dogs	0.065				
		building	0.058				
		pillow	0.047				
		refrigerator	0.047				
		furniture	0.047				
		pantry	0.047				

Comparing i-w2v to top-k, the i-w2v method outperforms the top-k in all vocabularies. In the High vocabulary, performance of the event *Rock Climbing* in top-k is slightly lower compared to the other direct matches, because in top-k the first occuring direct match is used instead of all direct matches. Using all direct matches for this event would improve MAP performance in All to 0.136.

Interestingly, CN outperforms both i-w2v and manual in the events *RenovateHome* in LowMid and All and *PlayingFetch* in LowMid. Table VIII shows the concepts and similarities of the different methods for the event *RenovateHome* in LowMid. In the event *PlayingFetch* in LowMid the addition of concepts, such as *throwing*, *ball* and *stick* (manual), decreases performance compared to only using the concept *dog* (CN).

6. DISCUSSION

Regarding the Vocabulary challenge, the results of the experiments show that a combination of multiple datasets improves performance. Although state of the art already tends to add as many datasets as possible to their vocabulary, we show that including high level concepts is important in video event retrieval. The results on the Vocabulary challenge show that using only the High vocabulary is better than using the LowMid vocabulary. The All vocabulary with both LowMid and High is also better than the LowMid. The correlation graph in Figure 2 shows that All is in the middle between LowMid and High. This observation makes us wonder if a combination of a LowMid and High vocabulary is indeed a good way to go, or if we should focus on a High vocabulary with more concepts. On one hand, the LowMid concepts are useful when no close matches of the High level concepts are present. On the other hand, the High level concepts can capture more than the combination of the LowMid level concepts. A related point is whether the high-level concepts can improve performance on lower level concept queries, such as *horse riding*. Will the high-level concept *horse riding competition*, possibly together with other events that include horse riding, improve performance on this query? In our opinion a concept on the same level of complexity as the query will provide the best performance, i.e. the query *horse riding* will achieve a higher retrieval performance with the matching concept *horse riding* compared to the concept *horse riding competition*, assuming both concept detectors perform accurately. In this example, the higher-level concept *horse riding competition* only includes a limited set of the query, resulting in a high precision but low recall situation. A lower-level concept, such as *horse* would include a set that is too broad, resulting in a high recall and low precision situation.

Regarding i-w2v, performance is better than current state of the art zero shot methods without re-training or re-ranking. I-w2v can be combined to the event pooling method from Lu et al. [2016] and the DCC method of Chang et al. [2016] to gain additional performance gain. The increase in performance compared to top-k does not seem significant, but when increasing the amount of concepts, the possibility of query drift is high. Current top-k strategy is to add only the most relevant concept. With a direct or near direct match between the event and the concepts, this is a reasonable strategy. In other tasks or with other events, this strategy is not optimal and a different number of k should be taken. Instead of optimizing the number k for each task, our strategy does not need this optimization. I-w2v is also able to combine concepts which cover different facets of the event, whereas other methods might only use the raw cosine similarity. Additionally, i-w2v does not seem that sensitive to the cutoff point, as shown in Table II.

Our proposed i-w2v method approaches the manual method. An advantage of the manual method is that human knowledge is richer than the knowledge in current knowledge bases or in word2vec, but the disadvantage is that 1) it requires a human to interpret all queries, which seems unfeasible in real-world applications; 2) it is hard for a human to indicate the proper weight. CN and w2v can automatically assign weights, but these weights are based on textual similarity. W2v learns from the context in which words appear, but the context does not indicate if the words are similar because they have an antonym (cat vs. dog), hyponym (chihuhua vs. dog), hypernym (animal vs. dog) or other type of relation. Knowledge bases such as ConceptNet have such relations, but for events little or no information is present. Because word2vec works as a vector model, the combination of multiple words in an user query gives better results than a combination of the different words searched in one of the knowledge bases. The method can, however, still be improved, because concepts with one directly matching word, such as *tree* in the concept *tree frog* for the event *FellingTree* and *home* in *home theater* for the event *RenovateHome*, sometimes retrieve a similarity that can be argued to be too high. But our word2vec method does not suffer from query drift and it approaches human performance, especially in a vocabulary that contains high-level concepts. In future work, an option could be to combine our method with the manual method by use of relevance feedback or use a hybrid method containing i-w2v and a knowledge base.

7. CONCLUSION

In this paper, we presented our Semantic Event Retrieval System that 1) includes high-level concepts and 2) uses a novel method in Concept Selection (*i-w2v*) based on semantic embeddings. Our experiments on the international TRECVID Multimedia Event Detection benchmark show that a vocabulary including high-level concepts can improve performance on the retrieval of complex and generic high-level events in videos, indicating the importance of high-level concepts in a vocabulary. Second, we show that our proposed Concept Selection method outperforms state of the art.

ACKNOWLEDGMENTS

We would like to thank the technology program Making Sense of Big Data (MSoBD) for their financial support. The work described in this paper was supported in part by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 120213).

REFERENCES

Robin Aly, Djoerd Hiemstra, Franciska de Jong, and Peter MG Apers. 2012. Simulating the future of concept-based video retrieval under improved detector performance. *Multimedia Tools and Applications* 60, 1 (2012), 203–231.

Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. 2011. Event detection and recognition for semantic annotation of video. *Multimedia Tools and Applications* 51, 1 (2011), pp. 279–302.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)* 44, 1 (2012), pp. 1.

Xiaojun Chang, Yi Yang, Alexander G Hauptmann, Eric P Xing, and Yao-Liang Yu. 2015. Semantic concept discovery for large-scale zero-shot event detection. In *Proceedings of the 24th International Conference on Artificial Intelligence*. AAAI Press, 2234–2240.

Xiaojun Chang, Yi Yang, Guodong Long, Chengqi Zhang, and Alexander G Hauptmann. 2016. Dynamic concept composition for zero-example event detection. *arXiv preprint arXiv:1601.03679* (2016).

Jiawei Chen, Yin Cui, Guangnan Ye, Dong Liu, and Shih-Fu Chang. 2014. Event-driven semantic concept discovery by exploiting weakly tagged internet images. In *Proc. of Int. Conf. on Multimedia Retrieval*. ACM, 1.

Jeffrey Dalton, James Allan, and Pranav Mirajkar. 2013. Zero-shot video retrieval using content and concepts. In *Proc. of the 22nd ACM Int. Conf. Information & Knowledge Management*. ACM, 1857–1860.

Maaike de Boer, Klamer Schutte, and Wessel Kraaij. 2015. Knowledge based query expansion in complex multimedia event detection. *Multimedia Tools and Applications* (2015), 1–19.

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

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

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

Amirhossein Habibian, Koen EA van de Sande, and Cees GM Snoek. 2013. Recommendations for video event recognition using concept vocabularies. In *Proc. of the 3rd Int. Conf. on Multimedia Retrieval*. ACM, 89–96.

Alexander Hauptmann, Rong Yan, and Wei-Hao Lin. 2007a. How many high-level concepts will fill the semantic gap in news video retrieval?. In *Proceedings of the 6th ACM international conference on Image and video retrieval*. ACM, 627–634.

Alexander Hauptmann, Rong Yan, Wei-Hao Lin, Michael Christel, and Howard Wactlar. 2007b. Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE transactions on multimedia* 9, 5 (2007), 958–966.

Bouke Huurnink, Katja Hofmann, and Maarten De Rijke. 2008. Assessing concept selection for video retrieval. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*. ACM, 459–466.

Mihir Jain, Jan C van Gemert, Thomas Mensink, and Cees GM Snoek. 2015. Objects2action: Classifying and localizing actions without any video example. In *Proc. of the IEEE Int. Conf. on Computer Vision*. 4588–4596.

Lu Jiang, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. 2014a. Easy samples first: Self-paced reranking for zero-example multimedia search. In *Proc. of the ACM Int. Conf. on Multimedia*. ACM, 547–556.

Lu Jiang, Teruko Mitamura, Shoou-I Yu, and Alexander G Hauptmann. 2014b. Zero-example event search using multimodal pseudo relevance feedback. In *Proc. of Int. Conf. on Multimedia Retrieval*. ACM, 297.

Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G Hauptmann. 2015b. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *ACM Int. Conf. on Multimedia Retrieval*. 27–34.

Yu-Gang Jiang, Subhabrata Bhattacharya, Shih-Fu Chang, and Mubarak Shah. 2012. High-level event recognition in unconstrained videos. *Int. Journal of Multimedia Information Retrieval* (2012), 1–29.

Yu-Gang Jiang, Zuxuan Wu, Jun Wang, Xiangyang Xue, and Shi-Fu Chang. 2015a. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. In *arXiv preprint arXiv:1502.07209*.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale Video Classification with Convolutional Neural Networks. In *CVPR*. 1725–1732.

Lyndon Kennedy and Alexander Hauptmann. 2006. LSCOM lexicon definitions and annotations (version 1.0). (2006).

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*. 2177–2185.

Omer Levy, Yoav Goldberg, and Ido Dagan. 2015. Improving distributional similarity with lessons learned from word embeddings. *Trans. of the Association for Computational Linguistics* 3 (2015), 211–225.

Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. 2007. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition* 40, 1 (2007), 262–282.

Yi-Jie Lu, Hao Zhang, Maaike de Boer, and Chong-Wah Ngo. 2016. Event Detection with Zero Example: Select the Right and Suppress the Wrong Concepts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. ACM, 127–134.

Masoud Mazloom, Efstratios Gavves, Koen van de Sande, and Cees Snoek. 2013. Searching informative concept banks for video event detection. In *Proc. of the 3rd Int. Conf. on Multimedia Retrieval*. ACM, 255–262.

Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. 2014. COSTA: Co-occurrence statistics for zero-shot classification. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2441–2448.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), pp. 39–41.

David Milne and Ian H Witten. 2013. An open-source toolkit for mining Wikipedia. *Artificial Intelligence* 194 (2013), pp. 222–239.

Apostol Paul Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. 2007. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proc. of the 15th Int. Conf. on Multimedia*. ACM, 991–1000.

Shi-Yong Neo, Jin Zhao, Min-Yen Kan, and Tat-Seng Chua. 2006. Video retrieval using high level features: Exploiting query matching and confidence-based weighting. In *International Conference on Image and Video Retrieval*. Springer, 143–152.

Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, and Georges Quenot. 2014. TRECVID 2014 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. of TRECVID 2014*. NIST, USA.

Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, Georges Quenot, and Roeland Ordelman. 2015. TRECVID 2015 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics. In *Proc. of TRECVID 2015*. NIST, USA.

Pushpa B Patil and Manesh B Kokare. 2011. Relevance Feedback in Content Based Image Retrieval: A Review. *Journal of Applied Computer Science & Mathematics* 10, 10 (2011), pp. 40–47.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

Alan F Smeaton, Paul Over, and Wessel Kraaij. 2006. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*. ACM, 321–330.

Steve Spagnola and Carl Lagoze. 2011. Edge dependent pathway scoring for calculating semantic similarity in ConceptNet. In *Proc. of the Ninth Int. Conf. on Computational Semantics*. Association for Computational Linguistics, 385–389.

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

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2014. Learning Spatiotemporal Features with 3D Convolutional Networks. In *arXiv preprint arXiv:1412.0767*.

Christos Tzelepis, Damianos Galanopoulos, Vasileios Mezaris, and Ioannis Patras. 2015. Learning to detect video events from zero or very few video examples. *Image and Vision Computing* (2015).

Shuang Wu, Sravanthi Bondugula, Florian Luisier, Xiaodan Zhuang, and Prem Natarajan. 2014. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2665–2672.

Shicheng Xu, Huan Li, Xiaojun Chang, Shoou-I Yu, Xingzhong Du, Xuanchong Li, Lu Jiang, Zexi Mao, Zhenzhong Lan, Susanne Burger, and others. 2015. Incremental Multimodal Query Construction for Video Search. In *Proc. of the 5th ACM on Int. Conf. on Multimedia Retrieval*. ACM, 675–678.

Yan Yan, Yi Yang, Haoquan Shen, Deyu Meng, Gaowen Liu, Alex Hauptmann, and Nicu Sebe. 2015. Complex event detection via event oriented dictionary learning. In *Twenty-Ninth AAAI Conf. on Artificial Intelligence*.

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

Shoou-I Yu, Lu Jiang, and Alexander Hauptmann. 2014. Instructional videos for unsupervised harvesting and learning of action examples. In *Proc. of the ACM Int. Conf. on Multimedia*. ACM, 825–828.

Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. 2014. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*. 487–495.