# Bridging the Ultimate Semantic Gap: A Semantic Search Engine for Internet Videos

Lu Jiang[1], Shoou-I Yu[1], Deyu Meng[2], Teruko Mitamura[1], Alexander G. Hauptmann[1]
[1] School of Computer Science, Carnegie Mellon University
[2] School of Mathematics and Statistics, Xi'an Jiaotong University
{lujiang, iyu, teruko, alex}@cs.cmu.edu, dymeng@mail.xjtu.edu.cn

## ABSTRACT

Semantic search in video is a novel and challenging problem in information and multimedia retrieval. Existing solutions are mainly limited to text matching, in which the query words are matched against the textual metadata generated by users. This paper presents a state-of-the-art system for event search without any textual metadata or example videos. The system relies on substantial video content understanding and allows for semantic search over a large collection of videos. The novelty and practicality is demonstrated by the evaluation in NIST TRECVID 2014, where the proposed system achieves the best performance. We share our observations and lessons in building such a state-of-the-art system, which may be instrumental in guiding the design of the future system for semantic search in video.

## Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Search process; I.2.10 [**Vision and Scene Understanding**]: Video analysis

## General Terms

Algorithms, Experimentation, Performance

## Keywords

Semantic Search; Video Search; Video Understanding; 0Ex; Content-based Retrieval; Multimedia Event Detection

## 1. INTRODUCTION

The explosion of multimedia data is creating impacts on many aspects of society. The huge volumes of accumulated video data bring challenges for effective multimedia search. Existing solutions, such as shown in YouTube, are mainly limited to text matching where the query words are matched against the textual metadata generated by the uploader [7]. This solution, though simple, proves to be futile when such metadata are either missing or less relevant to the video content. Content-based search, on the other hand, searches semantic features such as people, scenes, objects and actions that are automatically detected in the video content. A representative content-based retrieval task, initiated by the TRECVID community, is called Multimedia Event Detection (MED) [32]. The task is to detect the occurrence of a main event in a video clip without any textual metadata. The events of interest are mostly daily activities ranging from "birthday party" to "changing a vehicle tire".

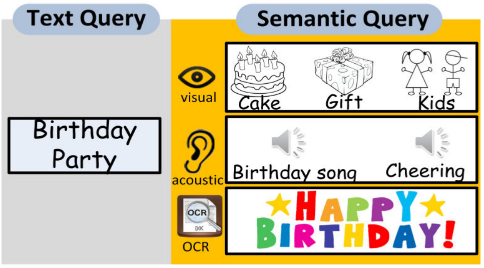


**Figure 1: Comparison of text and semantic query for the event "birthday party". Semantic queries contain visual/audio concepts about the video content.**

This paper discusses a setting in MED called zero-example search [18], or 0Ex for short, in which no relevant video is given in the query. 0Ex supports semantic search which relies on substantial video content understanding as opposed to shallow text matching in conventional methods. The query words are semantic features that are expected to occur in the relevant videos. For example, as illustrated in Fig. 1, the semantic query for the event "birthday party" might consist of visual concepts "cake", "gift" and "kids", audio concepts "birthday song" and "cheering sound". 0Ex also allows for flexible semantic search such as temporal or Boolean logic search. For example, searching for videos where opening gifts happens before consuming birthday cakes.

This paper details a state-of-the-art 0Ex system called E-Lamp semantic search engine, and according to National Institute of Standards and Technology (NIST), the proposed system achieves the best performance in TRECVID 2014 on a large collection of 200,000 Internet videos [32]. Our performance is about three times of the second best system. The outstanding performance is attributed to our rational pipeline as well as its effective components. We share our observations and lessons in building such a state-of-the-art system. The lessons are valuable because of not only the effort in designing and conducting numerous experiments but also the considerable computational resource to make the experiments possible. For example, building the semantic detectors costs us more than 1.2 million CPU core hours, which is equivalent to 140 years if it is running on a single core. We believe the shared lessons may significantly

*ICMR'15,* June 23–26, 2015, Shanghai, China.
Copyright © 2015 ACM 978-1-4503-3274-3/15/06 ...$15.00.
http://dx.doi.org/10.1145/2671188.2749399.

save the time and computational cycles for others who are interested in this problem.

The outstanding performance evaluated by NIST is a convincing demonstration of the novelty and practicality of the proposed system. Specifically, the novelty of this paper includes the solutions in system design and the insight on a number of empirical studies on semantic video search. The discussed techniques may also benefit other related tasks such as video summarization and recommendation. In summary, the contribution of this paper is twofold:

- We share our observations and lessons in building a state-of-the-art 0Ex event search system.
- Our pilot studies provide compelling insights on the comparison of modality contributions, semantic mapping methods and retrieval models for event search.

## 2. RELATED WORK

Multimedia event detection is an interesting problem. A number of studies have been proposed to tackle this problem on using several training examples (typically 10 or 100 examples) [14, 9, 38, 11, 31, 19, 34, 3, 36]. Generally, in a state-of-the-art system, the event classifiers are trained by low-level and high-level features, and the final decision is derived from the fusion of the individual classification results. For example, Habibian et al. [11] found several interesting observations about training classifiers only by semantic concept features. Gkalelis et al. [9] learned a representation for linear SVMs by subclass discriminant analysis, which yields 1-2 orders of magnitude speed-up. Wang et al. [38] discussed a notable system in TRECVID 2012 that is characterized by applying feature selection over so-called motion relativity features. Oh et al. [31] presented a latent SVM event detector that enables for temporal evidence localization. Jiang et al. [19] presented an efficient method to learn "optimal" spatial event representations from data.

Event detection with zero training examples is called 0Ex. It mostly resembles a real-world video search scenario, where users usually start the search without any example video. 0Ex is an understudied problem, and only few studies have been proposed very recently [6, 10, 26, 40, 18, 23]. Dalton et al. [6] discussed a query expansion approach for concept and text retrieval. Habibian et al. [10] proposed to index videos by composite concepts that are trained by combining the labeled data of individual concepts. Wu et al. [40] introduced a multimodal fusion method for semantic concepts and text features. Given a set of tagged videos, Mazloom et al. [26] discussed a retrieval approach to propagate the tags to unlabeled videos for event detection. Jiang et al. [18, 15] studied pseudo relevance feedback approaches which manage to significantly improve the original retrieval results. Existing related works inspire our system. However, to the best of our knowledge, there have been no studies on the 0Ex system architecture nor the analysis of each component.

## 3. FRAMEWORK

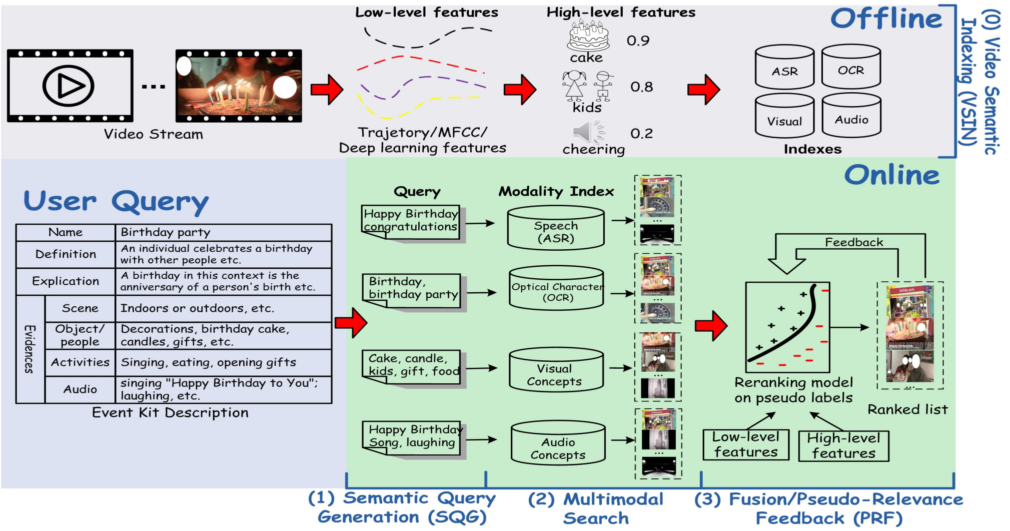
Semantic search in video can be modeled as a typical retrieval problem in which given a user query, we are interested in returning a ranked list of relevant videos. The proposed system comprises four major components, namely Video Semantic INdexing (VSIN), Semantic Query Generation (SQG), Multimodal Search, and Pseudo-Relevance Feedback (PRF)/Fusion, where VSIN is an offline indexing component, and the rest are the online search modules.

The VSIN component extracts semantic features from input videos, and indexes them for efficient online search. Typically, a video clip is first represented by low-level features such as dense trajectory features [39] for visual modality or deep learning features [27, 28] for audio modality. The low-level features are then input into the off-the-shelf detectors to extract the high-level features. Each dimension of the high-level feature corresponds to a confidence score of detecting a semantic concept in the video [14, 13]. Compared with low-level features, high-level features have a much lower dimension, which makes them economic for both storage and computation. The visual/audio concepts, Automatic Speech Recognition (ASR) [27] and Optical Character Recognition (OCR) are four types of high-level features in the system, in which ASR and OCR are textual features. ASR provides complementary information for events that are characterized by acoustic evidence. It especially benefits close-to-camera and narrative events such as "town hall meeting" and "asking for directions". OCR captures the characters in videos with low recall but high precision. The recognized characters are often not meaningful words but sometimes can be a clue for fine-grained detection, e.g. distinguishing "baby shower" and "wedding shower". The union of the dictionary vocabulary of high-level features constitutes the *system vocabulary*.

Users can express a query in various forms, such as a few concept names, a sentence or a structured description. NIST provides a query in the form of event-kit descriptions, which includes a name, definition, explication and visual/acoustic evidences (see the left corner of Fig. 2). The SQG component translates a user query into a multimodal *system query*, all words of which exist in the *system vocabulary*. Since the vocabulary is usually limited, addressing the out-of-vocabulary issue is a major challenge for SQG. The mapping between the user and system query is usually achieved with the aid of an ontology such as WordNet and Wikipedia. For example, a user query "golden retriever" may be translated to its most relevant alternative "large-sized dog", as the original concept may not exist in the system vocabulary.

Given a system query, the multimodal search component aims at retrieving a ranked list for each modality. As a pilot study, we are interested in leveraging the well-studied text retrieval models for video retrieval. To adapt the difference between semantic features and text features, we empirically study a number of classical retrieval models on various types of semantic features. We then apply the model to its most appropriate modalities. One notably benefit of doing so is that it can easily leverage the existing infrastructures and algorithms originally designed for text retrieval.

PRF (also known as reranking) refines a ranked list by reranking its videos. A generic PRF method first selects a few feedback videos, and assign assumed positive or negative labels to them. Since no ground-truth label is used, the assumed labels are called pseudo labels. The pseudo samples are then used to build a reranking model to improve the original ranked lists. A recent study shows that reranking can be modeled as a self-paced learning process [15], where the reranking model is built iteratively from easy to more complex samples. The easy samples are the videos ranked at the top, which are generally more relevant than those ranked lower. In addition to PRF, the standard late fusion is applied in our system.



| | | |
|---|---|---|
| Name | | Birthday party |
| Definition | | An individual celebrates a birthday with other people etc. |
| Explication | | A birthday in this context is the anniversary of a person's birth etc. |
| Evidences | Scene | Indoors or outdoors, etc. |
| | Object/ people | Decorations, birthday cake, candles, gifts, etc. |
| | Activities | Singing, eating, opening gifts |
| | Audio | singing "Happy Birthday to You"; laughing, etc. |

Event Kit Description

**Figure 2: Overview of the E-Lamp video search system.**

# 4. SYSTEM IMPLEMENTATIONS

The devil is in the detail. Careful implementations often turn out to be the cornerstone in many systems. To this end, this section discusses each component in more detail.

## 4.1 Large-scale Semantic Indexing

Concept detectors can be trained on still images or videos. The latter is more desirable due to the minimal domain difference and the capability for action and audio detection. A visual/audio concept in our system is represented as a multimodal document which includes a name, description, category, reliability (accuracy) and examples of the top detected video snippet. This definition provides a more tangible understanding about the concept for users.

The quantity (relevance) and quality of the semantic detectors are two crucial factors in affecting performance. The relevance is measured by the coverage of the concept vocabulary to the query, and thus is query-dependent. For convenience, we name it quantity as a larger vocabulary tends to increase the coverage. Quality is evaluated by the accuracy of the detector. Given limited resources, there exists a tradeoff between quality and quantity, i.e. building many unreliable detectors versus building a few reliable detectors. This tradeoff has been studied when there are several training examples [12, 11]. This paper presents a novel understanding about the tradeoff when there is no training example. The observations suggest that training more reasonably accurate detectors tends to be a sensible strategy.

Training detectors on large-scale video datasets increases both quantity and quality. But it turns out to be quite challenging. We approach this challenging problem with effort in two aspects. In theoretical aspect, a novel and effective method named self-paced curriculum learning [17] is explored. The idea is that, as opposed to training a detector using all samples at a time, we train increasingly complex detectors iteratively on balanced data subsets. The scheme of selecting samples to be trained in each iteration is controlled by a regularizer [22, 16, 17], and can be conveniently replaced to fit various problems. As for practical aspect, the module is optimized by storing kernel matrices in large shared-memory machines. This strategy yields 8 times speedup in training, enabling us training 3 thousand detectors over around 2 million videos (and video segments). In practice, in order to index huge volume of video data, the detectors need to be linear models (linear SVM or logistic regression), and nonlinear models have to be first transformed to the linear models (e.g. by explicit feature mapping [37]).

The ASR module is built on Kaldi [33] by training the HMM/GMM acoustic model with speaker adaptive training [27, 29] on videos. The trigram language model is pruned aggressively to speed up decoding. OCR is extracted by a commercial toolkit. The high-level features are indexed. The confidence scores for ASR and OCR are discarded, and the words are indexed by the standard inverted index. The visual/audio concepts are indexed by dense matrices to preserve their detection scores.

## 4.2 Semantic Query Generation

SQG translates a user query into a multimodal *system query* which only contains words in the system vocabulary. The first step in SQG is to parse negations in the query to recognize counter-examples. The recognized examples can be either discarded or to be associated with a “NOT” operator in the system query. Given TRECVID provides user queries in the form of event-kit description (see the left corner of Fig. 2), an event can be represented by the event name (1-3 words) or the frequent words in the event-kit description (after removing the template and stop words). These representations can be directly used as system queries for ASR/OCR as their vocabularies are sufficiently large to cover most words. For visual/audio concepts, the representations are used to map the out-of-vocabulary query words to their most relevant concepts in the system vocabulary. This mapping in SQG is challenging because of the complex relation between concepts. The relation between concepts includes mutual exclusion, subsumption, and frequent co-occurrence. For example, “cloud” and “sky” are frequently co-occurring concepts; “dog” subsumes “terrier”; and “blank frame” excludes “dog”. Our system includes the following mapping algorithms to map a word in the user query to the concept in the system vocabulary:

**Exact word matching**: A straightforward mapping is matching the exact query word (usually after stemming) against the concept name or description. Generally, for unambiguous words, it has high precision but low recall.

**WordNet mapping**: This mapping calculates the similarity between two words in terms of their distance in the

WordNet taxonomy. The distance can be defined in various ways such as structural depths in the hierarchy [41] or shared overlaps between synonymous words [1]. WordNet mapping is good at capturing synonyms and subsumption relations between two nouns.

**PMI mapping**: The mapping calculates the Point-wise Mutual Information (PMI) [5] between two words. Suppose $q_i$ and $q_j$ are two words in a user query, we have:

$$\mathrm{pmi}(q_i; q_j) = \log \frac{P(q_i, q_j|C_{ont})}{P(q_i|C_{ont})P(q_j|C_{ont})}, \tag{1}$$

where $P(q_i|C_{ont}), P(q_j|C_{ont})$ represent the probability of observing $q_i$ and $q_j$ in the ontology $C_{ont}$ (e.g. a collection of Wikipedia articles), which is calculated by the fraction of the document containing the word. $P(q_i, q_j|C_{ont})$ is the probability of observing the document in which $q_i$ and $q_j$ both occur. PMI mapping assumes that similar words tend to cooccur more frequently, and is good at capturing frequently co-occurring concepts (both nouns and verbs).

**Word embedding mapping**: This mapping learns a word embedding that helps predict the surrounding words in a sentence [30, 24]. The learned embedding, usually by neural network models, is in a lower-dimensional vector space, and the cosine coefficient between two words is often used to measure their distance. It is fast and also able to capture the frequent co-occurred words in similar contexts.

## 4.3 Multimodal Search

Given a system query, the multimodal search component aims at retrieving a ranked list for each modality. We are interested in leveraging the well-studied text retrieval models for video retrieval. There is no single retrieval model that can work the best for all modalities. As a result, our system incorporates several classical retrieval models and applies them to their most appropriate modalities. Let $Q = q_1, \ldots, q_n$ denote a system query. A retrieval model ranks videos by the score $\mathrm{s}(d|Q)$, where $d$ is a video in the video collection $C$. Our system includes the following retrieval models:

**Vector Space Model (VSM)**: This model represents both a video and a query as a vector of the words in the system vocabulary. The common vector representation includes generic term frequency (tf) and term frequency-inverse document frequency (tf-idf) [42]. $\mathrm{s}(d|Q)$ derives from either the product or the cosine coefficient between the video and the query vector.

**Okapi BM25**: This model extends tf-idf representation:

$$\mathrm{s}(d|Q)=\sum_{i=1}^{n}\log\frac{|C| - df(q_i)+\frac{1}{2}}{df(q_i)+\frac{1}{2}}\frac{tf(q_i, d)(k_1+1)}{tf(q_i, d)+k_1(1-b+b\frac{len(d)}{\overline{len}})}, \tag{2}$$

where $|C|$ is the total number of videos. $df(\cdot)$ returns the document frequency for a given word in the collection; $tf(q_i, d)$ returns the raw term frequency for the word $q_i$ in the video $d$. $len(d)$ calculates the sum of concept or word detection scores in the video $d$, and $\overline{len}$ is the average video length in the collection. $k_1$ and $b$ are two parameters to tune [25]. In the experiments, we set $b = 0.75$, and tune $k_1$ in $[1.2, 2.0]$.

**Language Model-JM Smoothing (LM-JM)**: The score is considered to be generated by a unigram language model [45]:

$$\mathrm{s}(d|Q) = \log P(d|Q) \propto \log P(d) + \sum_{i=1}^{n} \log P(q_i|d), \tag{3}$$

where $P(d)$ is usually assumed to be following the uniform distribution, i.e. the same for every video. $P(q_i|d)$ equals:

$$P(q_i|d) = \lambda \frac{tf(q_i, d)}{\sum_w tf(w, d)} + (1 - \lambda)P(q_i|C), \tag{4}$$

where $w$ enumerates all words or concepts in a given video, and $P(q_i|C)$ is called a smoother that can be calculated by $df(q_i)/|C|$. Eq. (4) linearly interpolates the maximum likelihood estimation (first term) with the collection model (second term) by a coefficient $\lambda$. The parameter is usually tuned in the range of $[0.7, 0.9]$. This model is good for retrieving long text queries, e.g. the frequent words in the event kit description.

**Language Model-Dirichlet Smoothing (LM-DL)**: This model adds a conjugate prior to the language model:

$$P(q_i|d) = \frac{tf(q_i, d) + \mu P(q_i|C)}{\sum_w tf(w, d) + \mu}, \tag{5}$$

where $\mu$ is the coefficient balancing the likelihood model and the conjugate prior, and is usually tuned in $[0, 2000]$ [45]. This model is good for short text queries, e.g. the event name representation.

## 4.4 Pseudo Relevance Feedback

PRF (or reranking) is a cost-effective method in improving performance. Studies have shown that it might hurt a few queries but generally improves the overall performance across all queries. We incorporate a general multimodal PRF method called SPaR [15] in our system.

**Self-Paced Reranking (SPaR)**: This method is inspired by the cognitive and learning process of humans and animals that gradually learning from easy to more complex samples [2, 22, 17, 46]. The easy samples in this problem are the videos ranked at the top, which are generally more relevant than those ranked lower. SPaR can take the input of either a single fused ranked list or a number of ranked lists from each of the modalities.[1] For convenience of notation, we mainly discuss the case of a single ranked list. Let $\Theta$ denote the reranking model. $\mathbf{x}_i$ denotes the feature of the $i$th video and $y_i$ for the $i$th pseudo label. Note that $x_i$ could be both high-level and low-level features [18]. We have:

$$\begin{aligned} &\min_{\Theta,\mathbf{y},\mathbf{v}} \sum_{i=1}^{n} v_i L(\mathbf{x}_i, y_i; \Theta) + f(\mathbf{v}; \lambda) \\ &\text{s.t. Constraints on } \Theta, \mathbf{y} \in \{-1, +1\}^n, \mathbf{v} \in [0, 1]^n, \end{aligned} \tag{6}$$

where $\mathbf{v} = [v_1, \ldots, v_n]$ are latent weight variables for each sample; $L$ is the loss function in $\Theta$; $f$ is self-paced function which determines the learning scheme, i.e. how to select and weight samples. $\lambda$ is a parameter controlling the model age. When $\lambda$ is small, a few easy samples with smaller loss will be considered. As $\lambda$ grows, more samples with larger loss will be gradually appended to train a mature model.

To run PRF in practice, we first need to pick a reranking model $\Theta$ (e.g. SVM or regression model), a self-paced function (e.g. binary, linear or mixture weighting) [15, 17], and reasonable starting values for the pseudo labels. The starting values can be initialized either by top ranked videos in the retrieved ranked lists or by other PRF methods. After the initialization, we iterate the following three steps:

[1] When different retrieval models are used for different features, and their ranked lists have very different score distributions, we may need to solve a linear programming problem to determine the starting pseudo positives [18].

1) training a model based on the selected pseudo samples and their weights (fixing $\mathbf{v}, \mathbf{y}$, optimize $\Theta$); 2) calculating pseudo positive samples and their weights by the self-paced function $f$, and selecting some pseudo negative samples randomly (fixing $\Theta$, optimize $\mathbf{v}, \mathbf{y}$)[2]; the weights of pseudo positive samples may be directly calculated by the closed-form solutions of $f$ [15]; 3) increasing the model age to include more positive samples in the next iteration (increase $\lambda$). The value of $\lambda$ is increased not by the absolute value but by the number of positive samples to be included. For example, if 5 samples need to be included, we can set $\lambda$ by the $6th$ smallest loss so that the samples whose loss are greater than $\lambda$ will have 0 weight. In our system, MMPRF [18] and SPaR [15] are incorporated, in which MMPRF is used to assign the starting values, and SPaR is used as the core algorithm. Average fusion of the PRF result with and original ranked list is used to obtain better results.

PRF is a cost-effective method in improving the performance of event search. We observed that the precision of the pseudo positives determines the performance of PRF (see Fig.5 in [18]). Therefore, high precision features such as OCR and ASR often turns out to be very useful. In addition, soft weighting self-paced functions that discriminate samples at different ranked positions tend to be better than binary weighting functions [15]. In this case, the model should support sample weighting (e.g. Libsvm tools-weights [4]). We observed two scenarios where the discussed reranking method could fail. First, SPaR may not help when the accuracy of the starting pseudo positive samples is below some threshold (e.g. the precision of pseudo positive is less than 0.1). This may be due to less relevant queries or poor quality of the high-level features. In this case, SPaR may not be useful. Second, SPaR may not help when the features used in PRF are not discriminative.

# 5. EXPERIMENTS

## 5.1 Setups

**Dataset and evaluation**: The experiments are conducted on the TRECVID Multimedia Event Detection (MED) MED13Test and MED14Test, evaluated by the official metric Mean Average Precision (MAP). Each set includes 20 events and 25,000 testing videos. The official test split released by NIST is used, and the reported MAP is comparable with others on the same split. The experiments are conducted without using any examples. We also evaluate each experiment on 10 randomly generated splits to reduce the bias brought by the split partition. The mean and 90% confidence interval are reported. Besides, the official results on MED14Eval evaluated by NIST TRECVID is also reported in the performance overview.

**Features and queries**: High-level features are Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), and visual semantic concepts. ImageNet features are trained on still images by deep convolution neural networks [21]. The rest are directly trained on videos by the SVM-based self-paced learning pipeline [16, 17]. The video datasets include: Sports [20], Yahoo Flickr Creative Common (YFCC) [35], Internet Archive Creative Common (IACC) [32] and Do it Yourself (DIY) [43]. The details of these datasets can be found in Table 1. In total, 3,043 video-based concept detectors are trained using the improved dense trajectory features [39], which costs more than 1.2 million CPU core hours (1,024 cores for about 2 months) in Pittsburgh Super-computing Center. Once trained, the detection (semantic indexing) for test videos is very fast. Two types of low-level features are used: dense trajectories [39] and MFCC [44] in the PRF model. The detailed configuration about PRF is available in the supplementary materials[3]. The input user query is the event-kit description. The system query is obtained by a two-step procedure: a preliminary mapping is automatically generated by the discussed mapping algorithms. The results are then examined by human experts to figure out the final system query. See supplementary materials for the example of user and system queries. For ASR/OCR, the automatically generated event name and description representation are used as the system query. Note that manual query examination is allowed in TRECVID and is used by many teams [32]. The automatically generated queries will be discussed in Section 5.5.

## 5.2 Performance Overview

We first examine the overall MAP (×100) of the full system. Table 2 lists the MAP and the runtime (ms/query)[4] across six datasets. The average MAP and the 90% confidence interval on the 10-splits are reported on MED13Test and MED14Test. The ground-truth data on MED14Eval has never been released, and thus the MAP evaluated by NIST TRECVID is only available on the single split. We observed that the results on a single split can be deceiving as no training data makes the result prone to overfitting. Therefore, we evaluate the MAP also on the ten splits. We observed that an improvement is statistically significant if it can be observed in the both cases.

Our system achieves the best performance across all of the datasets. According to [32], for example, Fig. 3 illustrates the comparison on the largest set of around 200,000 Internet videos, where the $x$-axis lists the system of each team, and the $y$-axis denotes the MAP evaluated by NIST TRECVID. The red bars represent our system with and without PRF. As we see, our system achieves the best performance which is about three times of the second best system. The improvement of PRF is evident as it is the only difference between the two red bars in Fig. 3. It is worth noting that the evaluation is very rigid because each system can only submit a single run within 60 minutes after getting the query, and the ground-truth data is not released even after the submission. In addition, for the ad-hoc events (see Fig. 3(b)) the query are generated online and unknown to the system beforehand. Since blind queries are more challenging, all systems perform worse on ad-hoc events. Nevertheless, our system still achieves an outstanding MAP.

## 5.3 Modality/Feature Contribution

Table 3 compares the modality contribution, where each run represents a system with a certain configuration. The MAP is evaluated on MED14Test and MED13Test, where the ground-truth data is available. As we see, visual modality is the most contributing modality, which by itself can recover about 85% MAP of the full system. ASR and OCR

[2] The reason we can randomly select pseudo negative samples is that they have negligible impacts on performance [18].

[3] `http://www.cs.cmu.edu/~lujiang/0Ex/icmr15.html`

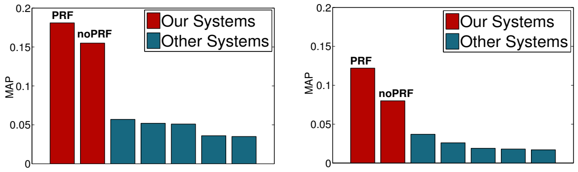
[4] The time is the search time evaluated on an Intel Xeon 2.53GHz CPU. It does not include the PRF time, which is about 238ms over 200K videos.

**Table 1: Summary of the datasets for training semantic visual concepts. ImageNet features are trained on still images, and the rest are trained on videos.**

| Dataset | #Samples | #Classes | Category | Example Concepts |
|---|---|---|---|---|
| DIY [43, 44] | 72,000 | 1,601 | Instructional videos | Yoga, Juggling, Cooking |
| IACC [32] | 600,000 | 346 | Internet archive videos | Baby, Outdoor, Sitting down |
| YFCC [35] | 800,000 | 609 | Amateur videos on Flickr | Beach, Snow, Dancing |
| ImageNet [8] | 1,000,000 | 1000 | Still images | Bee, Corkscrew, Cloak |
| Sports [20] | 1,100,000 | 487 | Sports videos on YouTube | Bullfighting, Cycling, Skiing |

**Table 2: Overview of the system performance.**

| Dataset | #Videos | MAP (×100) | | time |
|---|---|---|---|---|
| | | 1-split | 10-splits | (ms) |
| MED13Test | 25K | 20.75 | 19.47±1.19 | 80 |
| MED14Test | 25K | 20.60 | 17.27±1.82 | 112 |
| MED14EvalSub-PS | 32K | 24.1 | - | 192 |
| MED14EvalSub-AH | 32K | 22.1 | - | 200 |
| MED14Eval-PS | 200K | 18.1 | - | 1120 |
| MED14Eval-AH | 200K | 12.2 | - | 880 |



(a) Pre-Specified (PS) (b) Ad-Hoc (AH)

**Figure 3: The official results released by NIST TRECVID 2014 on MED14Eval (200, 000 videos).**

provide contribution to the full system but prove to be much worse than the visual features. PRF is beneficial in improving the performance of the full system.

To understand the feature contribution, we conduct leave-one-feature-out experiments. The performance drop, after removing the feature, can be used to estimate its contribution to the full system. As we see in Table 4, the results show that every feature provides some contribution. As the feature contribution is mainly dominated by a number of discriminative events, the comparison is more meaningful at the event-level (see supplementary materials[3]), where one can tell that, for example, the contribution of ImageNet mainly comes from three events E031, E015 and E037. Though the MAP drop varies on different events and datasets, the average drop on the two datasets follows: Sports > ImageNet > ASR > IACC > YFCC > DIY > OCR. The biggest contributor Sports is also the most computationally expensive feature. In fact, the above order of semantic concepts highly correlates to #samples in their datasets, which suggests the rationale of training concepts over big data sets.

**Table 3: Comparison of modality contribution.**

| Run | MED13Test | | MED14Test | |
|---|---|---|---|---|
| | 1-split | 10-splits | 1-split | 10-splits |
| FullSys+PRF | 22.12 | - | 22.05 | - |
| FullSys | 20.75 | 19.47±1.19 | 20.60 | 18.77±2.16 |
| VisualSys | 18.31 | 18.30±1.11 | 17.58 | 17.27±1.82 |
| ASRSys | 6.53 | 6.90±0.74 | 5.79 | 4.26±1.19 |
| OCRSys | 2.04 | 4.14±0.07 | 1.47 | 2.20±0.73 |

## 5.4 Quantity (Relevance) & Quality Tradeoff

To study the concept quantity (relevance) and quality trade-off, we conduct the following experiments where the IACC and ImageNet concepts are gradually added to the query, one at a time, according to their relevance. The relevance judgment is manually annotated by a group of assessors on 20 events. Fig. 4 illustrates the results on a representative event, where the $x$-axis denotes the number of concepts included, and the $y$-axis is the average precision. Each curve represents a trial of only selecting concepts of certain quality. The quality is manually evaluated by the precision of the top 10 detected examples on a third dataset. For example, the red curve indicates selecting only the relevant concepts whose precision of the top 10 detected examples is greater than 0.5. The blue curve marked by circles ($p@10 > 0$) is a trial of the largest vocabulary but on average the least accurate concepts. In contrast, the dashed curve is of the smallest vocabulary but the most accurate concepts, and thus for a query, there are fewer relevant concepts to choose from. As we see, neither of the settings is optimal. Selecting concepts with a reasonable precision, between 0.2 and 0.5 in this case, seems to be a better choice. The results suggest that incorporating more relevant concepts with reasonable quality is a sensible strategy. We also observed that merely increasing the number of low-quality concepts may not be helpful.

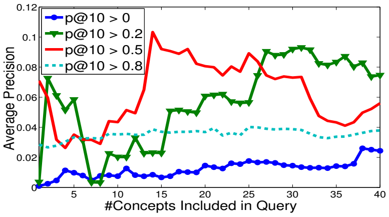


**Figure 4: Adding relevant concepts in "E037 Parking a vehicle" using concepts of different precisions.**

## 5.5 Semantic Matching in SQG

We apply the SQG algorithms to map the user query to the concept in the vocabulary. We use two metrics to compare these mapping algorithms. One is the precision of the 5 most relevant concepts returned by each algorithm. We manually assess the relevance for 10 events (E031-E040) on 4 concept features (i.e. 200 pairs for each mapping algorithm). The other is the MAP obtained by the 3 most relevant concepts. Table 5 lists the results, where the last column lists the runtime of calculating the mapping between 1,000 pairs of words. The second last row (Fusion) indicates the average fusion of the results of all mapping algorithms. As we see, in terms of P@5, PMI is slightly better than others, but it is also the slowest because its calculation involves looking up a index of 6 million articles in Wikipedia. Fusion of all mapping results yields a better P@5.

We then combine the automatically mapped semantic concepts with the automatically generated ASR and OCR query to test the MAP. Here we assume users have specified which feature to use for each query, and SQG is used to automatically find relevant concepts or words in the specified features. Our result can be understood as an overestimate of a fully-automatic SQG system, in which users do not even need to specify the feature. As we see in Table 5, PMI performs the best on MED13Test whereas on MED14Test it is Exact Word Matching. The fusion of all mapping results (the

Table 4: Comparison of feature contribution.

| SysID | Visual Concepts | | | | | ASR | OCR | MAP | | MAP Drop(%) |
|---|---|---|---|---|---|---|---|---|---|---|
| | IACC | Sports | YFCC | DIY | ImageNet | | | 1-split | 10-splits | |
| MED13/IACC | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 18.93 | 18.61±1.13 | 9% |
| MED13/Sports | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | 15.67 | 14.68±0.92 | 25% |
| MED13/YFCC | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | 18.14 | 18.47±1.21 | 13% |
| MED13/DIY | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 19.95 | 18.70±1.19 | 4% |
| MED13/ImageNet | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | 18.18 | 16.58±1.18 | 12% |
| MED13/ASR | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | 18.48 | 18.78±1.10 | 11% |
| MED13/OCR | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | 20.59 | 19.12±1.20 | 1% |
| MED14/IACC | | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 18.34 | 17.79±1.95 | 11% |
| MED14/Sports | ✓ | | ✓ | ✓ | ✓ | ✓ | ✓ | 13.93 | 12.47±1.93 | 32% |
| MED14/YFCC | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ | 20.05 | 18.55±2.13 | 3% |
| MED14/DIY | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | 20.40 | 18.42±2.22 | 1% |
| MED14/ImageNet | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | 16.37 | 15.21±1.91 | 20% |
| MED14/ASR | ✓ | ✓ | ✓ | ✓ | ✓ | | ✓ | 18.36 | 17.62±1.84 | 11% |
| MED14/OCR | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | | 20.43 | 18.86±2.20 | 1% |

second last row) improves the MAP on both the datasets. We then fine-tune the parameters of the mapping fusion and build our AutoSQG system (the last row). As we see, AutoSQG only achieves about 55% of the full system's MAP. Several reasons account for the performance drop: 1) the concept name does not accurately describe what is being detected; 2) the quality of mapping is limited (P@5=0.42); 3) relevant concepts may not necessarily be discriminative. For example, "animal" and "throwing ball" appear to be relevant to the query "playing a fetch", but the former is too general and the latter is about throwing a baseball which is visually different; "dog" is much less discriminative than "group of dogs" for the query "dog show". The results suggest that the automatic SQG is not well-understood. The proposed automatic mappings are still very preliminary, and could be further refined by manual inspection. We found it is beneficial to represent a concept as a multimodal document that includes a name, description, category, reliability (accuracy) and examples of the top detected video snippet.

Table 5: Comparison of SQG mapping algorithms.

| Mapping Method | P@5 | MAP | | Time (s) |
|---|---|---|---|---|
| | | 13Test | 14Test | |
| Exact Word Matching | 0.340 | 9.66 | 7.22 | 0.10 |
| WordNet | 0.330 | 7.86 | 6.68 | 1.22 |
| PMI | 0.355 | 9.84 | 6.95 | 22.20 |
| Word Embedding | 0.335 | 8.79 | 6.21 | 0.48 |
| Mapping Fusion | 0.420 | 10.22 | 9.38 | - |
| AutoSQGSys | - | 12.00 | 11.45 | - |

## 5.6 Comparison of Retrieval Methods

Table 6 compares the retrieval models on MED14Test using four features: ASR, OCR and two types of visual concepts. As we see, there is no single retrieval model that works the best for all features. For ASR and OCR words, BM25 and Language Model with JM smoothing (LM-JM) yield the best MAPs. An interesting observation is that VSM can only achieve 50% MAP of LM-JM on ASR (2.94 versus 5.79). This observation suggests that the role of retrieval models in video search is substantial. For semantic concepts, VSM performs no worse than other models. We hypothesize that it is because the concept representation is dense, i.e. every dimension has a nonzero value, and thus is quite different from sparse text features. To verify this hypothesis, we sparsify the Sports representation using a basic approach, where we only preserve the top $m$ dimensions in a video, and $m$ is set proportional to the concept vocabulary size. As we see, the sparse feature with 1% nonzero elements can recover 85% of its dense representation. Besides, BM25 and LM exhibit better MAPs in the sparse representation. Since the dense representation is difficult to index and search, the results suggest a promising direction for large-scale search using the sparse semantic feature.

Table 6: Comparison of retrieval models on MED14Test using ASR, OCR, Sports and IACC.

| Feat. | Split | VSM-tf | VSM-tfidf | BM25 | LM-JM | LM-DP |
|---|---|---|---|---|---|---|
| ASR | 1 | 2.94 | 1.26 | 3.43 | **5.79** | 1.45 |
| | 10 | 2.67 | 1.49 | 3.03 | **4.26** | 1.14 |
| OCR | 1 | 0.56 | 0.47 | **1.47** | 1.02 | 1.22 |
| | 10 | 2.50 | 2.38 | **4.52** | 3.80 | 4.07 |
| Sports | 1 | **9.21** | 8.97 | 8.83 | 8.75 | 7.57 |
| | 10 | **10.61** | 10.58 | 10.13 | 10.25 | 9.04 |
| IACC | 1 | 3.49 | **3.52** | 2.44 | 2.96 | 2.06 |
| | 10 | **2.88** | 2.77 | 2.05 | 2.45 | 2.08 |

Table 7: Study of retrieval performance using sparse concept features (Sports) on MED14Test.

| Density | VSM-tf | BM25 | LM-JM | LM-DP |
|---|---|---|---|---|
| 1% | 9.06 | **9.58** | 9.09 | 9.38 |
| 2% | 9.93 | 10.12 | **10.14** | 10.07 |
| 4% | 10.34 | 10.36 | 10.26 | **10.38** |
| 16% | **10.60** | 10.45 | 10.03 | 9.89 |
| 100% | **10.61** | 10.13 | 10.25 | 9.04 |

# 6. CONCLUSIONS AND FUTURE WORK

We proposed a state-of-the-art semantic video search engine called E-Lamp. The proposed system goes beyond conventional text matching approaches, and allows for semantic search without any textual metadata or example videos. We shared our lessons on system design and compelling insights on a number of empirical studies. From the experimental results, we arrive at the following recommendations.

- **Recommendation 1**: Training concept detectors on big data sets is ideal. However, given limited resources, building more detectors of reasonable accuracy seems to be a sensible strategy. Merely increasing the number of low-quality concepts may not improve performance.
- **Recommendation 2**: PRF (or reranking) is an effective approach to improve the search result.
- **Recommendation 3**: Retrieval models may have substantial impacts to the search result. A reasonable strategy is to incorporate multiple models and apply them to their appropriate features/modalities.
- **Recommendation 4**: Automatic query generation for queries in the form of event-kit descriptions is still very challenging. Combining mapping results from various mapping algorithms and applying manual examination afterward is the best strategy known so far.

This paper is merely a first effort towards semantic search in Internet videos. The proposed system can be improved in various ways, e.g. by incorporating more accurate visual and audio concept detectors, by studying more appropriate retrieval models, by exploring search interfaces or interactive search schemes. As shown in our experiments, the automatic semantic query generation is not well understood. Closing the gap between the manual and automatic query may point to a promising direction.

## Acknowledgments

The authors would like to thank Yajie Miao, Chuang Gan, Shicheng Xu, and Susanne Burger for meaningful discussions and kind support.

This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. Deyu Meng was partially supported by the China NSFC project under contract 61373114. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. It used the Blacklight system at the Pittsburgh Supercomputing Center (PSC).

## 7. REFERENCES

[1] S. Banerjee and T. Pedersen. An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing*, 2002.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[3] S. Bhattacharya, F. X. Yu, and S.-F. Chang. Minimally needed evidence for complex event recognition in unconstrained videos. In *ICMR*, 2014.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2:27:1–27:27, 2011.

[5] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29, 1990.

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

[7] J. Davidson, B. Liebald, J. Liu, et al. The youtube video recommendation system. In *RecSys*, 2010.

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

[9] N. Gkalelis and V. Mezaris. Video event detection using generalized subclass discriminant analysis and linear support vector machines. In *ICMR*, 2014.

[10] A. Habibian, T. Mensink, and C. G. Snoek. Composite concept discovery for zero-shot video event detection. In *ICMR*, 2014.

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

[12] A. Hauptmann, R. Yan, W.-H. Lin, M. Christel, and H. Wactlar. Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. *TMM*, 9(5):958–966, 2007.

[13] N. Inoue and K. Shinoda. n-gram models for video semantic indexing. In *MM*, 2014.

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

[15] L. Jiang, D. Meng, T. Mitamura, and A. G. Hauptmann. Easy samples first: Self-paced reranking for zero-example multimedia search. In *MM*, 2014.

[16] L. Jiang, D. Meng, S.-I. Yu, Z. Lan, S. Shan, and A. G. Hauptmann. Self-paced learning with diversity. In *NIPS*, 2014.

[17] L. Jiang, D. Meng, Q. Zhao, S. Shan, and A. G. Hauptmann. Self-paced curriculum learning. In *AAAI*, 2015.

[18] L. Jiang, T. Mitamura, S.-I. Yu, and A. G. Hauptmann. Zero-example event search using multimodal pseudo relevance feedback. In *ICMR*, 2014.

[19] L. Jiang, W. Tong, D. Meng, and A. G. Hauptmann. Towards efficient learning of optimal spatial bag-of-words representations. In *ICMR*, 2014.

[20] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] M. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *NIPS*, 2010.

[23] H. Lee. Analyzing complex events and human actions in” in-the-wild” videos. In *UMD Ph.D Theses and Dissertations*, 2014.

[24] O. Levy and Y. Goldberg. Dependency-based word embeddings. In *ACL*, 2014.

[25] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[26] M. Mazloom, X. Li, and C. G. Snoek. Few-example video event retrieval using tag propagation. In *ICMR*, 2014.

[27] Y. Miao, L. Jiang, H. Zhang, and F. Metze. Improvements to speaker adaptive training of deep neural networks. In *SLT*, 2014.

[28] Y. Miao and F. Metze. Improving low-resource cd-dnn-hmm using dropout and multilingual dnn training. In *INTERSPEECH*, 2013.

[29] Y. Miao, F. Metze, and S. Rawat. Deep maxout networks for low-resource speech recognition. In *ASRU*, 2013.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, 2013.

[31] S. Oh, S. McCloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. A. Perera, M. Pandey, and J. J. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Machine vision and applications*, 25(1):49–69, 2014.

[32] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, W. Kraaij, A. F. Smeaton, and G. Quéenot. TRECVID 2014 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *TRECVID*, 2014.

[33] D. Povey, A. Ghoshal, G. Boulianne, et al. The kaldi speech recognition toolkit. In *ASRU*, 2011.

[34] B. Safadi, M. Sahuguet, and B. Huet. When textual and visual information join forces for multimedia retrieval. In *ICMR*, 2014.

[35] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li. The new data and new challenges in multimedia research. *arXiv preprint arXiv:1503.01817*, 2015.

[36] W. Tong, Y. Yang, L. Jiang, et al. E-lamp: integration of innovative ideas for multimedia event detection. *Machine Vision and Applications*, 25(1):5–15, 2014.

[37] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. *PAMI*, 34(3):480–492, 2012.

[38] F. Wang, Z. Sun, Y. Jiang, and C. Ngo. Video event detection using motion relativity and feature selection. In *TMM*, 2013.

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

[40] S. Wu, S. Bondugula, F. Luisier, X. Zhuang, and P. Natarajan. Zero-shot event detection using multi-modal fusion of weakly supervised concepts. In *CVPR*, 2014.

[41] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *ACL*, 1994.

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

[43] S.-I. Yu, L. Jiang, and A. Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *MM*, 2014.

[44] S.-I. Yu, L. Jiang, Z. Xu, et al. Cmu-informedia@trecvid 2014. In *TRECVID*, 2014.

[45] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *TOIS*, 22(2), 2004.

[46] Q. Zhao, D. Meng, L. Jiang, Q. Xie, Z. Xu, and A. G. Hauptmann. Self-paced learning for matrix factorization. In *AAAI*, 2015.