Fast and Accurate Content-based Semantic Search in 100M Internet Videos

Lu Jiang[1], Shoou-I Yu[1], Deyu Meng[2],
Yi Yang[3], Teruko Mitamura[1], Alexander G. Hauptmann[1]
[1] Carnegie Mellon University, [2] Xi'an Jiaotong University, [3] University of Technology Sydney
{lujiang, iyu, teruko, alex}@cs.cmu.edu, dymeng@mail.xjtu.edu.cn

ABSTRACT

Large-scale content-based semantic search in video is an interesting and fundamental problem in multimedia analysis and retrieval. Existing methods index a video by the raw concept detection score that is dense and inconsistent, and thus cannot scale to “big data” that are readily available on the Internet. This paper proposes a scalable solution. The key is a novel step called concept adjustment that represents a video by a few salient and consistent concepts that can be efficiently indexed by the modified inverted index. The proposed adjustment model relies on a concise optimization framework with interpretations. The proposed index leverages the text-based inverted index for video retrieval. Experimental results validate the efficacy and the efficiency of the proposed method. The results show that our method can scale up the semantic search while maintaining state-of-the-art search performance. Specifically, the proposed method (with reranking) achieves the best result on the challenging TRECVID Multimedia Event Detection (MED) zero-example task. It only takes 0.2 second on a single CPU core to search a collection of 100 million Internet videos.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; I.2.10 [**Vision and Scene Understanding**]: Video analysis

General Terms

Algorithms, Experimentation, Performance

Keywords

Internet Video Search; Semantic Search; Big Data; Content-based Retrieval; Multimedia Event Detection; Zero Example

1. INTRODUCTION

Searching semantic content in Internet videos has long been a goal of multimedia analysis and retrieval. This fundamental problem is a building block for many tasks such as video visualization, recommendation and summarization. As opposed to merely searching user-generated metadata, such as titles and descriptions, content-based semantic search strives to leverage concepts that are automatically detected in the video, such as objects/scenes/actions. For example, in order to search for videos depicting a “birthday party”, we might look for visual concepts like “cake”, “gift” and “kids”, and audio concepts like “birthday songs” and “cheering sounds”. This semantic search relies on extensive video understanding, and requires neither metadata nor example videos. According to the National Institute of Standards and Technology (NIST), semantic search in video is also known as zero-example search (0Ex). A benchmark task called Multimedia Event Detection (MED) 0Ex, which was initiated by NIST TRECVID in 2013, is to detect the occurrence of a main event, e.g. “making a sandwich” or “rock climbing”, occurring in a video clip without any user-generated metadata or example videos.

A number of studies have demonstrated promising progress in this direction [17, 9, 22, 42, 14, 21, 1]. However, existing methods index a video by the raw concept detection score that is dense and inconsistent. This solution is mainly designed for analysis and search over a few thousand of videos, and cannot scale to big data collections required for real world applications. For example, in order to search the semantics in YouTube videos, a system must be able to search over billions of Internet videos [35]. The scale problem is well beyond the scope of the existing work, and thus, as shown in the experiments in Section 7, their solutions will simply fail as the data grows. Large-scale semantic search, though challenging, opens possibilities for many interesting applications. For example, a currently nonexisting functionality is to search videos on social media platforms such as Facebook or Twitter. 12 million videos are posted on Twitter every day that have either no text or only a few words with little relevance to the visual content. It is extremely difficult to find meaningful information without content-based semantic search. Another example relates to in-video advertising. Currently, it may be hard for companies to effectively place in-video advertisements as the user-generated metadata typically does not describe the video content, let alone concept occurrences in time. However, a solution may be achieved by putting the advertisement into the top-ranked relevant videos returned by the semantic search. For example, a sport shoe company may use the query “(running OR jumping) AND urban_scene AND parkour” to select parkour videos for special shoe ads.

MM'15, October 26–30, 2015, Brisbane, Australia.
© 2015 ACM. ISBN 978-1-4503-3459-4/15/10 ...$15.00.
DOI: http://dx.doi.org/10.1145/2733373.2806237.

Even though a modern text retrieval system can already search over billions of text documents, the task is still very challenging for semantic video search. The main reason is that semantic concepts are quite different from the text words, and indexing of semantic concepts is still an understudied problem. Specifically, concepts are automatically extracted by detectors with limited accuracy. The raw detection score associated with each concept is inappropriate for indexing for two reasons. First, the distribution of the scores is dense, i.e. a video contains every concept with a non-zero detection score, which is analogous to a text document containing every word in the English vocabulary. The dense score distribution hinders effective inverted indexing and search. Second, the raw score may not capture the complex relations between concepts, e.g. a video may have a "puppy" but not a "dog". This type of inconsistency can lead to inaccurate search results.

To address this problem, we propose a novel step called concept adjustment that aims at producing video (and video shot) representations that tend to be consistent with the underlying concept representation. After adjustment, a video is represented by a few salient and consistent concepts that can be efficiently indexed by the inverted index. In theory, the proposed adjustment model is a general optimization framework that incorporates existing techniques as special cases. In practice, as demonstrated in our experiments, the adjustment increases the consistency with the ground-truth concept representation on the real world TRECVID dataset. Unlike text words, semantic concepts are associated with scores that indicate how confidently they are detected. We propose an extended inverted index structure that incorporates the real-valued detection scores and supports complex queries with Boolean and temporal operators.

To the best of our knowledge, our study is the first effort devoted to applying the semantic video search techniques [18] to million-scale search applications. Compared to existing methods, the proposed method exhibits the following three benefits. First, it advances the text retrieval method for video retrieval. Therefore, while existing methods fail as the size of the data grows, our method is scalable, extending the current capability of semantic search by a few orders of magnitude while maintaining state-of-the-art performance. The experiments in Section 6.2 and Section 7 validate this argument. Second, we propose a novel component called concept adjustment in a common optimization framework with solid probabilistic interpretations. Finally, our empirical studies shed some light on the tradeoff between efficiency and accuracy in a large-scale video search system. These observations will be helpful in guiding the design of future systems on related tasks such as video summarization, recommendation or hyperlinking [1].

The experimental results are promising on three datasets. On the TRECVID Multimedia Event Detection (MED), our method achieves comparable performance to state-of-the-art systems, while reducing its index by a relative 97%. The results on the TRECVID Semantic Indexing dataset demonstrate that the proposed adjustment model is able to generate more accurate concept representation than baseline methods. The results on the largest public multimedia dataset called YCCC100M [37] show that the method is capable of indexing and searching over a large-scale video collection of 100 million Internet videos. It only takes 0.2 seconds on a single CPU core to search a collection of 100 million Internet videos. Notably, the proposed method with reranking is able to achieve by far the best result on the TRECVID MED 0Ex task, one of the most representative and challenging tasks for semantic search in video. In summary, our contribution is threefold:

- We propose a scalable solution that extends the current capability of semantic video search by a few orders of magnitude of data while maintaining state-of-the-art accuracy.
- We propose a novel optimization framework that represents a video with relatively few salient and consistent concepts. Several related techniques can be regarded as its special cases.
- Our paper is the first work that addresses a long-lasting challenge of content-based semantic search in 100 million Internet videos on a single core.

2. RELATED WORK

Traditional content-based video retrieval methods have demonstrated promising results in a number of large scale applications, such as SIFT matching [34, 25] and near duplicate detection [43]. The search mainly utilizes the low-level descriptors that carry little semantic meaning. On the other hand, semantic video search aims at searching the high-level semantic concepts automatically detected in the video content. Compared with traditional methods, semantic search relies on understanding about the video content. This line of study first emerged in a TRECVID task called Semantic Indexing [27], the goal of which is to search the occurrence of a single or a pair of concepts [36]. A concept can be regarded as a visual or acoustic semantic tag on people, objects, scenes and actions in the video content [26].

With the advance in object and action detection, people started to focus on searching more complex queries called events. An event is more complex than a concept as it usually involves people engaged in process-driven actions with other people and/or objects at a specific place and time [13]. For example, the event "rock climbing" involves video clips such as outdoor bouldering, indoor artificial wall climbing or snow mountain climbing. A benchmark task on this topic is called TRECVID Multimedia Event Detection (MED). Its goal is to detect the occurrence of a main event occurring in a video clip without any user-generated metadata. MED is divided into two scenarios in terms of whether example videos are provided. When example videos are given, a state-of-the-art system first train classifiers using multiple features and fuse the decision of the individual classification results [7, 40, 11, 28, 32, 2, 39, 10, 3].

This paper focuses on the other scenario named zero-example search (0Ex) where no example videos are given. 0Ex mostly resembles a real world scenario, in which users start the search without any example. As opposed to training an event detector, 0Ex searches semantic concepts that are expected to occur in the relevant videos, e.g. we might look for concepts like "car", "bicycle", "hand" and "tire" for the event "changing a vehicle tire". A few studies have been proposed on this topic [9, 22, 42, 17, 14, 21]. A closely related work is detailed in [18], where the authors presented their lessons and observations in building a state-of-the-art semantic search engine for Internet videos. Existing solutions are promising but only for a few thousand videos because they cannot scale to big data collections. Therefore, the biggest collection in existing studies contains no more than 200 thousand videos [18, 29].

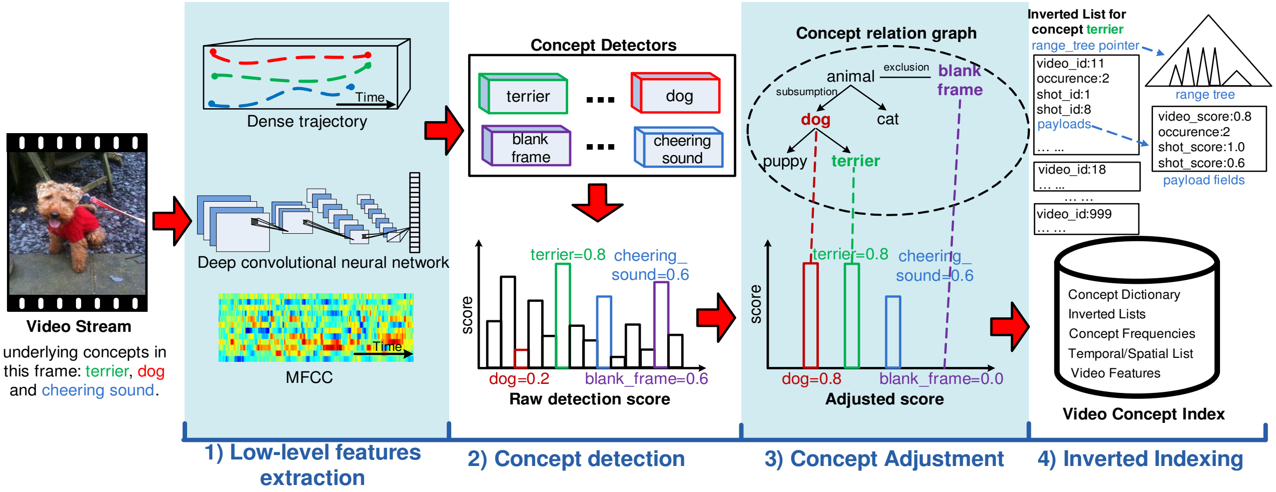


Figure 1: Pipeline for the offline video indexing.

3. OVERVIEW

There is an offline stage called video semantic indexing before one can perform any online search. The stage aims at indexing the semantic content in a video for efficient online search. As illustrated in Fig. 1, there are four major components in this video semantic indexing pipeline, namely, low-level feature extraction, concept detection, concept adjustment and inverted indexing, in which the concept adjustment component is first proposed in this paper.

A video clip is first represented by low-level visual or audio features. Common features include dense trajectories [41], deep learning [20] and MFCC features. The low-level features are then fed into the off-the-shelf detectors to extract the semantic concept features, where each dimension corresponds to a confidence score of detecting a semantic (audio and visual) concept in a video shot. The dimensionality is equal to the number of unique detectors in the system. The semantic concept is a type of high-level feature [13] that can be used in search. The high-level features also include Automatic Speech Recognition (ASR) [23, 24] and Optical Character Recognition (OCR) [45].

We found that raw concept detection scores are inappropriate for indexing for two reasons: *distributional inconsistency* and *logical inconsistency*. The *distributional inconsistency* means that the distribution of the raw detection score is inconsistent with the underlying concept distribution of the video. The underlying concept representation tends to be sparse but the distribution of the detection score is dense, i.e. a video contains every concept. Indexing the dense representation by either dense matrices or inverted indexes is known to be inefficient. For example, Fig. 1 illustrates an example in which the raw concept detection contains 14 nonzero scores but there are only three concepts in the underlying representation: “dog”, “terrier”, and “cheering sound”. As we see, the dense distribution of the raw detection score is very different from the underlying distribution.

The *logical inconsistency* means that the detection scores are not consistent with the semantic relation between concepts, e.g. a video contains a “terrier” but not a “dog”. This type of inconsistency results from that 1) the detectors are usually trained by different people using different data, features and models. It is less likely for them to consider the concept consistency that is not in their vocabulary; 2) even within a concept vocabulary, many classification models cannot capture complex relation between concepts [4]. The inconsistent representation can lead to inaccurate search results if not properly handled. For example, in Fig. 1, the score of “dog” 0.2 is less than the score of “terrier” 0.8; the frame is detected as “blank frame”, which means a empty frame, and a “terrier”.

To address the problem of distributional and logical inconsistencies, we propose a novel step called concept adjustment. It aims at generating consistent concept representations that can be efficiently indexed and searched. We propose an adjustment method based on the recently proposed label relation graph [4] that models the hierarchy and exclusion relation between concepts (see Step 3 in Fig. 1). After adjustment, a video is represented by a few salient concepts, which can be efficiently indexed by the inverted index. In addition, the adjusted representation is logically consistent with the complex relation between concepts.

An effective approach is to index the adjusted representation by the inverted index in text retrieval. However, unlike text words, semantic concepts are associated with scores that indicate how confidently they are detected. The detection score cannot be directly indexed by the standard inverted index. As a result, the scores are usually indexed by dense matrices in existing methods [42, 18]. To this end, we modify the inverted index structure so that it can index the real-valued adjusted score. The modified index contains inverted indexes, frequency lists that store the concept statistics used in the retrieval model, temporal lists that contain the shot information, and video feature lists that store the low-level features. The extended index structure is compatible to existing text retrieval algorithms. More details are given in Section 5.

4. CONCEPT ADJUSTMENT

4.1 Concept Adjustment Model

In this paper, we make no assumption on the training process of the off-the-shelf concept detectors. The detectors may be trained by any type of features, models or data. We relax the assumption [4] that detectors must be “re-trainable” by particular training algorithms because this is usually impossible when we do not have the access to the training data, the code or the computational resource.

Concept adjustment aims at generating video (or video shot) representations that tend to be consistent to the underlying concept representation and meanwhile can be searched efficiently. An ideal video representation tends to be similar to the underlying concept representation in terms of the distributional and logical consistency. To this end, we propose an optimization model to find consistent video representations of the given raw concept detection output. Formally, let $\mathbf{D} \in \mathbb{R}^{n \times m}$ denote the raw scores outputted by the concept detectors, where the row represents the n shots in a video, and the column represents the m visual/audio concepts. The prediction score of each concept is in the range between 0 and 1, i.e. $\forall i, j$, $\mathbf{D}_{ij} \in [0, 1]$. We are interested in obtaining a consistent representation $\mathbf{v} \in \mathbb{R}^{m \times 1}$, which can be obtained by solving the following optimization problem:

$$\begin{aligned} \mathop{\arg\min}_{\mathbf{v} \in [0,1]^m} & \frac{1}{2} \|\mathbf{v} - f_p(\mathbf{D})\|_2^2 + g(\mathbf{v}; \alpha, \beta) \\ & \text{subject to } \mathbf{A}\mathbf{v} \leq \mathbf{c} \end{aligned} \tag{1}$$

where

$$f_p(\mathbf{D}) = (1 - (\frac{m-1}{m})^p)[\|\mathbf{d}_1\|_p, \ldots, \|\mathbf{d}_m\|_p]^T \tag{2}$$

and $\mathbf{D} = \begin{bmatrix} | & & | \\ \mathbf{d}_1 & \cdots & \mathbf{d}_m \\ | & & | \end{bmatrix}$. Each element of $f_p(\mathbf{D})$ is the p-norm of the column vector of $\mathbf{D}$. $g(\mathbf{v}; \alpha, \beta)$ is a regularizer of $\mathbf{v}$ with the parameters α and β. $g(\cdot)$ imposes the distributional consistency, and will be discussed in Section 4.1.1. $\mathbf{A}$ and $\mathbf{c}$ are a constant matrix and a constant vector, which model the logical consistencies and will be discussed in Section 4.1.2. It is easy to verify that when $p = \infty$ and $p = 1$, the operator $f_p(\mathbf{D})$ corresponds to the max and the average pooling operator. Usually $g(\cdot)$ is convex, and thus Eq. (1) can be conveniently solved by the standard convex programming toolbox [8]. The raw prediction score may diminish during the concept adjustment. It is usually helpful to normalize the optimal value of $\mathbf{v} = [v_1, \cdots, v_m]$ by:

$$\hat{v}_i = \min(1, \frac{v_i}{\sum_{j=1}^m v_j} \sum_{j=1}^m f_p(\mathbf{D})_j I(v_j)), \tag{3}$$

where $I(v)$ is an indicator function equalling 1 when $v > 0$, and 0 otherwise. Here we define $0/0 = 0$.

In order to obtain the shot-level adjusted representation, we can treat a shot as a "video" and let $\mathbf{D}$ be a single row matrix containing the detection score of the shot. Eq. (1) can be used but with an extra integer set in the constraints (see Section 4.1.2).

4.1.1 Distributional Consistency

For the distributional consistency, a regularization term $g(\mathbf{v}; \alpha, \beta)$ is introduced that produces sparse representations while taking into account that certain concepts may co-occur together. A naive implementation is to use the l_0 norm:

$$g(\mathbf{v}; \alpha, \beta) = \frac{1}{2} \beta^2 \|v\|_0. \tag{4}$$

This regularization term presents a formidable computational challenge. In this paper we propose a more feasible and general regularization term. Suppose the concepts are divided into q non-overlapping groups. A group may contain a number of co-occurring concepts, or a single concept if it does not co-occur with others. Such sparsity and group sparsity information can be encoded into the model by adding a convex regularization term $g(\mathbf{v})$ of the l_1 norm and the sum of group-wise l_2 norm of $\mathbf{v}$:

$$g(\mathbf{v}; \alpha, \beta) = \alpha\beta\|\mathbf{v}\|_1 + (1-\alpha) \sum_{l=1}^q \beta\sqrt{p_l} \|\mathbf{v}^{(l)}\|_2, \tag{5}$$

where $\mathbf{v}^{(l)} \in \mathbb{R}^{p_l}$ is the coefficient for the lth group where p_l is the length of that group. $\alpha \in [0, 1]$ and β are two parameters controlling the magnitude of the sparsity.

The parameter α balances the group-wise and the within-group sparsity. When $\alpha = 1$, $g(\mathbf{v})$ becomes *lasso* [38] that finds a solution with few nonzero entries. When $\alpha = 0$, $g(\mathbf{v})$ becomes *group lasso* [46], that only yields nonzero entries in a sparse set of groups. If a group is included then all coefficients in the group will be nonzero. Sometimes, the sparsity within a group is also needed, i.e. if a group is included, only few coefficients in the group will be nonzero. This is known as *sparse-group lasso* [33] that linearly interpolates *lasso* and *group lasso* by the parameter α.

In the context of semantic concepts, *lasso* is an approximation to the corresponding l_0 norm regularization problem which is computationally expensive to solve. *Lasso* and the l_0 norm term assume the concepts are independent, and works well when the assumption is satisfied, e.g. the 1,000 concepts in ImageNet challenges where the concepts are manually selected to be exclusive labels [4]. On the other hand, *Group lasso* assumes the there exist groups of concepts that tend to be present or absent together frequently, e.g. "sky/cloud", "beach/ocean/waterfront" and "table/chair". The group may also include multimodal concepts such as "baby/baby noises". Since co-occurring concepts may not always be present together, the within-group sparse solution is needed sometimes, i.e. only few concepts in a group are nonzero. This can be satisfied by *sparse-group lasso* that makes weaker assumptions about the underlying concept distribution.

4.1.2 Logical Consistency

The concept relation is modeled by Hierarchy and Exclusion (HEX) graph. Deng et al. [4] recently introduced label relation graphs called Hierarchy and Exclusion (HEX) graphs. The idea is to infer a representation that maximizes the likelihood and do not violate the label relation defined in the HEX graph. Following Deng et al. [4], we assume that the graph is given beforehand. According to [4], a HEX graph is defined as:

Definition 1. *A HEX graph $G = (N, E_h, E_e)$ is a graph consisting of a set of nodes $N = \{n_1, \cdots, n_m\}$, directed edges $E_h \subseteq N \times N$ and undirected edges $E_e \subseteq N \times N$ such that the subgraph $G_h = (N, E_h)$ is a directed acyclic graph and the subgraph $G_e = (N, E_e)$ has no self-loop.*

Each node in the graph represents a distinct concept. A hierarchy edge $(n_i, n_j) \in E_h$ indicates that concept n_i subsumes concept n_j in the concept hierarchy, e.g. "dog" is a parent of "puppy". An exclusion edge $(n_i, n_j) \in E_e$ indicates concept n_i and n_j are mutually exclusive, e.g. a frame cannot be both "blank frame" and "dog". Based on Definition 1, we define the logically consistent representation as:

Definition 2. *$\mathbf{v} = [v_1, \cdots, v_m]$ is a vector of concept detection scores. The ith dimension corresponds to the concept node $n_i \in N$ in the HEX graph G. $\mathbf{v} \in [0, 1]^m$ is logically consistent with G if for any pair of concepts (n_i, n_j):*

1. if $n_i \in \alpha(n_j)$, then $v_i \geq v_j$;

2. if $\exists n_p \in \bar{\alpha}(n_i)$, $\exists n_q \in \bar{\alpha}(n_j)$ and $(n_p, n_q) \in E_e$, then we have $v_i v_j = 0$;

where $\alpha(n_i)$ is a set of all ancestors of n_i in G_h, and $\bar{\alpha}(n_i) = \alpha(n_i) \cup n_i$.

Definition 2 indicates that a logically consistent representation should not violate any concept relation defined in its HEX graph G. This definition generalizes the legal assignments in [4] to allow concepts taking real values. We model the logical consistency by the affine constraints $\mathbf{Av} \leq \mathbf{c}$. The constant matrix $\mathbf{A}$ and vector $\mathbf{c}$ can be calculated from Algorithm 1. For each edge in the graph, Algorithm 1 defines a constraint on values the two concepts can take. A hierarchy edge $(n_i, n_j) \in E_h$ means that the value of a parent is no less than the value of its children, e.g. "puppy=0.8" but "dog=0.2" is inconsistent. For each exclusion edge, Algorithm 1 introduces an affine constraint $v_i + v_j = 1$ and $v_i, v_j \in \{0, 1\}$ to avoid the case where two concepts both have nonzero values. Note that the solution of the exclusion constraint complies with the binary legal assignments in [4] that for any $(n_i, n_j) \in E_e$, $(v_i, v_j) \neq (1, 1)$. It is unnecessary to propagate an exclusion constraint to its children nodes because the hierarchy constraint guarantees the score of the children nodes is no more than their parent. According to Definition 2, it is easy to prove that the optimal solution of Eq. (1) is logically consistent with a given HEX graph. The problem with integer constraints can be solved either by the mixed-integer convex programming toolbox, or by the constraint relaxation [6].

THEOREM 1. *The optimal solutions of Eq.* (1) *(before or after normalization) is logically consistent with its given HEX graph.*

```
Algorithm 1: Constraints for logical consistency.
    input  : A HEX graph $G = (V, E_h, E_e)$
    output: A constant matrix $\mathbf{A}$ and a constant $\mathbf{c}$.
$n = |E_h| + |E_e|$; $m = |V|$; $k = 0$;
$\mathbf{A} = \mathbf{0}_{n \times m}$, $\mathbf{c} = \mathbf{0}_{n \times 1}$;
foreach $(n_i, n_j) \in E_h$ do
    $\mathbf{A}_{ki} = -1$; $\mathbf{A}_{kj} = 1$; $\mathbf{c}_k = 0$;
    $k$++;
end
Define an integer constraint set $\mathbb{I} \leftarrow \phi$;
foreach $(n_i, n_j) \in E_e$ do
    $\mathbf{A}_{ki} = 1$; $\mathbf{A}_{kj} = 1$; $\mathbf{c}_k = 1$;
    add $n_i$, $n_j$ to $\mathbb{I}$;
    $k$++;
end
return $\mathbf{A}$, $\mathbf{c}$, $\mathbb{I}$;
```

4.2 Discussions

The proposed model can produce a representation that tends to be both distributionally and logically consistent to the underlying concept representation. A nice property of the model in Eq. (1) is that it can degenerate to several existing methods. For example, it is easy to verify that the max and the average pooling results are optimal solutions of Eq. (1) in special cases. Theorem 1 indicates that the optimal solution of adjusted representations complies with the logical consistency definition. Theorem 2 indicates that the thresholding and the top-k thresholding results are optimal solutions of Eq. (1) in special cases. The thresholding method preserves scores only above some threshold. In some cases, instead of using an absolute threshold, one can alternatively set the threshold in terms of the number of concepts to be included. This is known as the top-k thresholding. The proof is provided in the supplementary materials.

THEOREM 2. *The thresholding and the top-k thresholding results are optimal solutions of Eq.* (1) *in special cases.*

The proposed model also provides common interpretations of what are being optimized. The physical meaning of the optimization problem in Eq. (1) can be interpreted as a maximum a priori model. The interpretation is provided in the supplementary materials.

The choice of the proposed model parameters depends on the underlying distribution of the semantic concepts. For the manually exclusive concepts, such as the 1,000 concepts in the ImageNet challenge [31], the l_0 norm or the l_1 norm without any HEX constraint should work reasonably well. In addition, as the model is simple, the problem can be efficiently solved by the closed-form solution. When the concepts are of concrete hierarchical or exclusion relations, such as the concepts in TRECVID SIN [29], incorporating the HEX constraint tends to be beneficial. The group-lasso and the sparse-group lasso play a role when groups of concepts tend to co-occur together frequently. It can be important for the multimodal concept detectors that capture the same concept by multiple features, e.g. audio or visual. An approach to derive the co-occurring concepts is by clustering the concepts in their labeled training data. We observed big clusters tend to include more loosely coupled concepts, e.g. sky/cloud is a good group, but sky/cloud/helicopter is not. To be prudent, we recommend limiting the group size in clustering.

Note the exclusion relation between concepts only makes sense at the shot-level adjustment, as in the video-level representation the scores of exclusive concepts can be both nonzeros. Solving a mixed integer convex programming problem takes more time than solving a regular convex programming problem. So when the proposed method is applied on shot-level features, it is useful to use some type of constraint relaxation techniques. Besides, in the current model, we assume the concept detectors are equally accurate. A simple extension to embed this information is by discounting the squared loss of inaccurate concepts in Eq. (1).

5. INVERTED INDEXING & SEARCH

The dense raw detection scores are usually indexed by dense matrices in existing methods [42, 18]. This simple solution, though preserves all detection scores, is not scalable. In comparison, the proposed adjustment method represents a video by a few salient and consistent concepts. To index the adjusted representation, we modify the structure of inverted index so that it can incorporate real-valued detection scores. In this section, we discuss video indexing and search, using the proposed inverted indexes.

5.1 Inverted Indexing

After adjustment, a video is represented by a few salient and consistent concepts. In analogy to words in a text document, concepts can be treated as "words" in a video. Unlike text words, concepts are associated with scores that indicate how confidently they are detected. The real-valued scores are difficult to be directly indexed in the standard inverted index designed for text words. A naive approach is by binning, where we assign real values to the bins representing the segment covering the numerical value. The concepts are duplicated by the number of its filled bins. However, this solution creates hallucinating concepts in the index, and cannot store the shot-level concept scores.

To solve the problem in a more principled way, we propose a modified inverted index to incorporate the real-valued detection scores. In text retrieval, each unique text word has a list of *postings* in the inverted index. A *posting* contains a document ID, the term frequency, and the term positions in the document. The term frequency is used in the retrieval model, and the position information is used in the proximity search. Inspired by this structure, in our system, the concept with a nonzero score in the adjusted representation is indexed to represent a video. Each unique concept has a list of *video postings* and a range search tree. An example index structure is illustrated in Step 4 in Fig. 1. A *video posting* contains a video ID, the number of concept occurrence in the video, a video-level detection score, and a list of video shots in which the concept occurs. It also has a payload to store the real-valued detection score for each shot. The query that searches for the video-level score of a certain range can be handled by the range tree, e.g. "videos that contain dog > 0.5"; the query that searches for the shot-level score can be handled by the payload in the posting, e.g. "shots that contain dog > 0.5"; otherwise, the query can be processed in a similar way as in text retrieval using the adjusted video-level score, e.g. videos that contain "dog AND cat".

5.2 Video Search

A search usually contains two steps: retrieving a list of *video postings* and ranking the postings according to some retrieval model. In our system, we consider the following query operators to retrieve a *video posting* list:

- **Modality query**: Searching a query term in a specified modality. For example, "visual:dog" returns the videos that contain the visual concept "dog"; "visual:dog/[score s_1, s_2]" returns the videos that have a detection score of "dog" between s_1 and s_2. "visual" is the default modality. The other modalities are "asr" for automatically recognized speech, "ocr" for recognized optical characters, and "audio" for audio concepts.
- **Temporal query**: Searching query terms that have constraints on their temporal occurrences in a video. The constraints can be specified in terms of the absolute timestamp like "videos that contain dog between the time t_1 and t_2", the relative sequence like "videos in which dog is seen before cat", or the proximity relations like "videos that contain dog and cat within the time window of t_1". A temporal query can be handled in a similar fashion as the proximity search in text retrieval.
- **Boolean query**: Multiple terms can be combined together with Boolean operators to form a more complex query. Our system supports three operators: "AND", "OR" and "AND NOT", where the "OR" operator is the default conjunction operator.

A Boolean query can be handled by the standard algorithms in text retrieval, as Theorem 1 guarantees that the adjusted representation is logically consistent. However, the query may be accelerated by utilizing the concept relation in the HEX graph. For example, it is unnecessary to run a query to realize that ("dog" AND "animal") = "dog". Suppose the query is expressed in the disjunctive normal form. Given a HEX graph G and two concepts $n_i, n_j \in V$, for each term in the disjunctive normal form, we apply: (n_i AND n_j) $= n_i$ if $n_j \in \alpha(n_i)$, where $\alpha(n_i)$ is the set of all ancestors of n_i in G_h; (n_i AND NOT n_j) $= \phi$ if $\exists n_p \in \bar{\alpha}(n_i)$, $\exists n_q \in \bar{\alpha}(n_j)$ and $(n_p, n_q) \in E_e$. The simplified query can be then used to retrieval the *video postings*.

After retrieving a *video posting* list, the next step is to rank the postings according to some retrieval model. A retrieval model can have substantial impact on the performance, and following [18], we adopt the Okapi BM25 model that work reasonably well for concept retrieval. Suppose the input query is $Q = q_1, \cdots, q_n$, the model ranks a video d by:

$$\mathrm{s}(d|Q) = \sum_{i=1}^{n} \log \frac{|C| - df(q_i) + \frac{1}{2}}{df(q_i) + \frac{1}{2}} \frac{tf(q_i, d)(k_1 + 1)}{tf(q_i, d) + k_1(1 - b + b\frac{len(d)}{\overline{len}})}, \tag{6}$$

where $|C|$ is the total number of videos. $tf(q_i, d)$ returns the score of the concept q_i in the adjusted representation of video d. $df(\cdot)$ calculates the sum of adjusted score of q_i in the video collection. $len(d)$ calculates the sum of adjusted scores for video d, and $\overline{len}$ is the average length across all videos. k_1 and b are two parameters to tune. Note the statistics are calculated by the adjusted concept score rather than the raw detection score.

6. EXPERIMENTS

6.1 Setups

Dataset and evaluation: The experiments are conducted on two TRECVID benchmarks called Multimedia Event Detection (MED): MED13Test and MED14Test [29]. The performance is evaluated by several metrics for a better understanding, which include: P@20, Mean Reciprocal Rank (MRR), Mean Average Precision (MAP), and MAP@20, where the MAP is the official metric used by NIST. Each set includes 20 events over 25,000 test videos. The official NIST's test split is used. We also evaluate each experiment on 10 randomly generated splits to reduce the split partition bias. All experiments are conducted without using any example or text metedata.

Features and queries: Videos are indexed by high-level features including semantic concepts, Automatic Speech Recognition (ASR), and Optical Character Recognition (OCR). These features are provided in [18]. For semantic concepts, 1,000 ImageNet concepts are trained by the deep convolution neural networks [20]. The remaining 3,000+ concepts are directly trained on videos by the self-paced learning pipeline [15, 16] on around 2 million videos using improved dense trajectories [41]. The video datasets include Sports [19], Yahoo Flickr Creative Common (YFCC100M) [37], Internet Archive Creative Common (IACC) [29] and Do it Yourself (DIY) [44]. The details of these datasets can be found in Table 1. The ASR module is built on Kaldi [30, 23]. OCR is extracted by a commercial toolkit. Three sets of queries are used: 1) *Expert* queries are obtained by human experts; 2) *Auto* queries are automatically generated by the Semantic Query Generation (SQG) methods in [18] using ASR, OCR and visual concepts; 3) *AutoVisual* queries are also automatically generated but only includes the visual concepts. The *Expert* queries are used by default.

Configurations: The concept relation released by NIST is used to build the HEX graph for IACC features [26][1]. The adjustment is conducted at the video-level average ($p = 1$ in Eq. (1)) so no shot-level exclusion relations are used. For other concept features, since there is no public concept relation specification, we manually create the HEX graph. The HEX graphs are empty for Sports and ImageNet features as there is no evident hierarchical and exclusion relation in their concepts. We cluster the concepts based on the correlation of their training labels, and include concepts that frequently co-occur together into a group. The parameters are tuned on a validation sets, and then are fixed across all experiment datasets including MED13Test, MED14Test and YFCC100M. Specifically, the default parameters in Eq. (1) are $p = 1$, $\alpha = 0.95$. β is set as the top k detection scores in a video, and is different for each type of features: 60 for IACC, 10 for Sports, 50 for YFCC100M, 15 for ImageNet, and 10 for DIY features. CVX optimization toolbox [8] is used to solve the model in Eq. (1). Eq. (6) is used as the retrieval model for concept features, where $k_1 = 1.2$ and $b = 0.75$.

6.2 Performance on MED

We first examine the overall performance of the proposed method. Table 2 lists the evaluation metrics over the two benchmarks on the standard NIST split and on the 10 randomly generated splits. The performance is reported over three set of queries: *Expert*, *Auto*, and *AutoVisual*.

Table 3 compares the performance of the raw and the adjusted representation on the 10 splits of MED13Test. *Raw* lists the performance of indexing the raw score by dense matrices; *Adjusted* lists the performance of indexing the adjusted concepts by the proposed index which preserves the real-valued scores. As we see, although *Raw* is slightly better than *Adjusted*, its index in the form of dense matrices is more than 33 times bigger than the inverted index in *Adjusted*. The comparison substantiates that the adjusted representation has comparable performances with the raw representation but can be indexed by a much smaller index.

An interesting observation is that *Adjusted* outperforms *Raw* on 8 out of 20 events on MED13Test (see the supplementary materials). We inspected the results and found that concept adjustment can generate more consistent representations. Fig. 2 illustrates raw and adjusted concepts on three example videos. Since the raw score is dense, we only list the top ranked concepts. As we see, the noisy concept in the raw detection may be removed by the logical consistency, e.g. "snow" in the first video. The missed concept may be recalled by logical consistencies, e.g. "vehicle" in the third video is recalled by "ground vehicle". The frequently co-occurring concepts may also be recovered by distributional consistencies, e.g. "cloud" and "sky" in the second video. Besides, we also found that Boolean queries can boost the performance. For example, in "E029: Winning a race without a vehicle", the query of relevant concepts such as swimming, racing or marathon can achieve an AP of 12.5. However, the Boolean query also containing "AND NOT" concepts such as car racing or horse riding can achieve an AP of 24.5.

Table 2: Overview of the system performance.
(a) Performance on the NIST's split

Dataset	Query	Evaluation Metric			
		P@20	MRR	MAP@20	MAP
MED13Test	Expert	0.355	0.693	0.280	0.183
	Auto	0.243	0.601	0.177	0.118
	AutoVisual	0.125	0.270	0.067	0.074
MED14Test	Expert	0.228	0.585	0.147	0.172
	Auto	0.150	0.431	0.102	0.100
	AutoVisual	0.120	0.372	0.067	0.086

(b) Average Performance on the 10 splits

Dataset	Query	Evaluation Metric			
		P@20	MRR	MAP@20	MAP
MED13Test	Expert	0.325	0.689	0.247	0.172
	Auto	0.253	0.592	0.187	0.120
	AutoVisual	0.126	0.252	0.069	0.074
MED14Test	Expert	0.219	0.540	0.144	0.171
	Auto	0.148	0.417	0.084	0.102
	AutoVisual	0.117	0.350	0.063	0.084

Table 3: Comparison of the raw and the adjusted representation on the 10 splits.

Method	Index	Evaluation Metric			
		P@20	MRR	MAP@20	MAP
MED13 Raw	385M	0.312	0.728	0.230	0.176
MED13 Adjusted	11.6M	0.325	0.689	0.247	0.172
MED14 Raw	357M	0.233	0.610	0.155	0.185
MED14 Adjusted	12M	0.219	0.540	0.144	0.171

We then compare our best result with the published results on MED13Test. The experiments are all conducted on the NIST's split, and thus are comparable to each other. As we see in Table 4, the proposed method has a comparable performance to the state-of-the-art methods. Notably, the proposed method with one iteration of reranking [14] is able to achieve the best result. The comparison substantiates that our method maintains state-of-the-art accuracy. It is worth emphasizing that the baseline methods may not scale to big data sets, as the dense matrices are used to index all raw detection scores [42, 14, 18].

Table 4: MAP (× 100) comparison with the published results on MED13Test.

Method	Year	MAP
Composite Concepts [9]	2014	6.4
Tag Propagation [22]	2014	9.6
MMPRF [17]	2014	10.1
Clauses [21]	2014	11.2
Multi-modal Fusion [42]	2014	12.6
SPaR [14]	2014	12.9
E-Lamp FullSys [18]	2015	20.7
Our System	2015	18.3
Our System + reranking	**2015**	**20.8**

The parameters α and β in Eq. (1) control the magnitude of sparsity in the concept adjustment, i.e. the percentage of concepts with nonzero scores in a video representation. A sparse representation reduces the size of indexes but hurts the performance at the same time. As we will see later, β is more important than α in affecting the performance. Therefore, we fix α to 0.95 and study the impact of β. Fig. 3 illustrates the tradeoff between accuracy and efficiency on the 10 splits of MED13Test. By tuning β, we obtain different percentages of nonzero concepts in a video representation. The x-axis lists the percentage in the log scale. $x = 0$ indicates the performance of ASR and OCR without semantic concept features. We discovered that we do not need many concepts to index a video, and a few adjusted concepts already preserve significant amount of information for search. As we see, the best tradeoff in this problem is 4% of the total

[1] http://www-nlpir.nist.gov/projects/tv2012/tv11.sin.relations.txt

Table 1: Summary of the semantic concept training sets. ImageNet features are trained on still images, and the rest are trained on videos.

Dataset	#Samples	#Classes	Category	Example Concepts
DIY [44]	72,000	1,601	Instructional videos	Yoga, Juggling, Cooking
IACC [29]	600,000	346	Internet archive videos	Baby, Outdoor, Sitting down
YFCC100M [37]	800,000	609	Amateur videos on Flickr	Beach, Snow, Dancing
ImageNet [5]	1,000,000	1000	Still images	Bee, Corkscrew, Cloak
Sports [19]	1,100,000	487	Sports videos on YouTube	Bullfighting, Cycling, Skiing

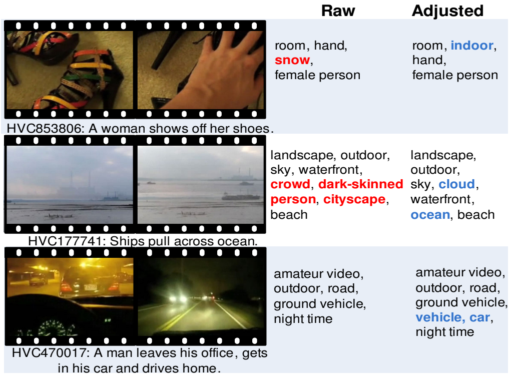

Raw
Adjusted
room, hand, snow, female person
room, indoor, hand, female person
HVC853806: A woman shows off her shoes.
landscape, outdoor, sky, waterfront, crowd, dark-skinned person, cityscape, beach
landscape, outdoor, sky, cloud, waterfront, ocean, beach
HVC177741: Ships pull across ocean.
amateur video, outdoor, road, ground vehicle, night time
amateur video, outdoor, road, ground vehicle, vehicle, car, night time
HVC470017: A man leaves his office, gets in his car and drives home.

Figure 2: Comparison of raw and adjusted concepts.

concepts (i.e. 163 concepts). Further increasing the number of concepts only leads to marginal performance gain.

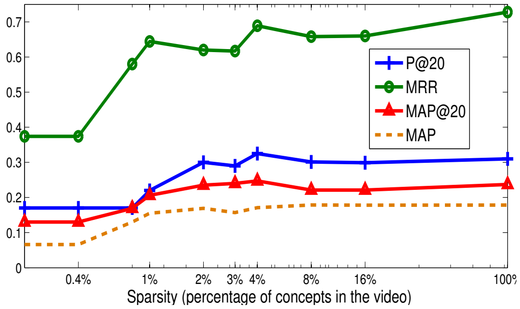


Figure 3: The impact of parameter β. $x = 0$ indicates the performance of ASR and OCR without semantic concepts.

6.3 Accuracy of Concept Adjustment

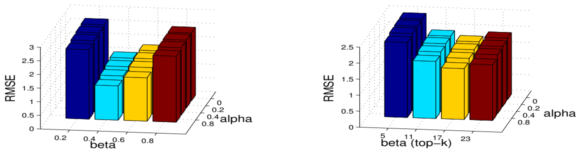
Generally the comparison in terms of retrieval performance depends on the query words. A query-independent way to verify the accuracy of the adjusted concept representation is by comparing it to the ground truth representation. To this end, we conduct experiments on the TRECVID Semantic Indexing (SIN) IACC set, where the manually labeled concepts are available for each shot in a video. We use our detectors to extract the raw shot-level detection score, and then apply the adjustment methods in Section 4 to obtain the adjusted representation. The performance is evaluated by Root Mean Squared Error (RMSE) to the ground truth concepts for the 1,500 test shots in 961 videos.

We compare our adjustment method with the baseline methods in Table 5, where HEX Graph indicates the logical consistent representation [4] on the raw detection scores (i.e. $\beta = 0$), and Group Lasso denotes the representation yield by Eq. (5) when $\alpha = 0$. We tune the parameter in each baseline method and report its best performance. As the ground truth label is binary, we let the adjusted scores be binary in all methods. As we see, the proposed method outperforms all baseline methods. We hypothesize the reason is that our method is the only one that combines the distributional consistency and the logical consistency. As discussed in Section 4.2, the baseline methods can be regarded as special cases of the proposed model.

We study the parameter sensitivity in the proposed model. Fig. 4 plots the RMSE under different parameter settings. Physically, α interpolates the group-wise and within-group sparsity, and β determines the number of concepts in a video. As we see, the parameter β is more sensitive than α, and accordingly we fix the value of α in practice. Note the parameter β is also an important parameter in the baseline methods including thresholding and top-k thresholding.

Table 5: Comparison of the adjusted representation and baseline methods on the TRECVID SIN set. The metric is Root Mean Squared Error (RMSE).

Method	RMSE
Raw Score	7.671
HEX Graph Only	8.090
Thresholding	1.349
Top-k Thresholding	1.624
Group Lasso	1.570
Our method	**1.236**



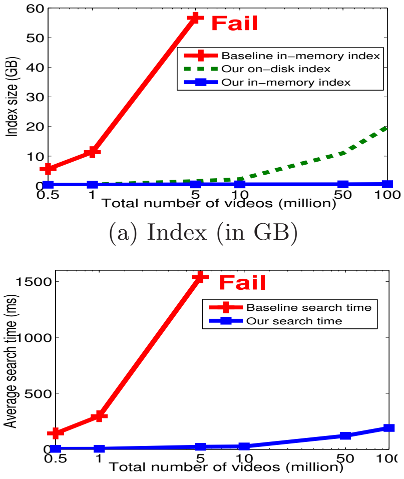
(a) Thresholding (b) Top-k thresholding

Figure 4: Sensitivity study on the parameter α and β in our model.

7. APPLICATION ON YFCC100M

We apply the proposed method on YFCC100M, the largest public multimedia collection that has ever been released [37]. It contains about 0.8 million Internet videos (approximately 12 million key shots) on Flickr. For each video and video shot, we extract the improved dense trajectory, and detect 3,000+ concepts by the off-the-shelf detectors in Table 1. We implement our inverted index based on Lucene [12], and a similar configuration described in 6.1 is used except we set $b = 0$ in the BM25 model. All experiments are conducted without using any example or text metedata. It is worth emphasizing that as the dataset is very big. The offline video indexing process costs considerable amount of computational resources in Pittsburgh super-computing center. To this end, we share this valuable benchmark with our community `http://www.cs.cmu.edu/~lujiang/0Ex/mm15.html`.

To validate the efficiency and scalability, we duplicate the original videos and video shots, and create an artificial set of 100 million videos. We compare the search performance of the proposed method to a common approach in existing studies that indexes the video by dense matrices [42, 18]. The experiments are conducted on a single core of Intel Xeon 2.53GHz CPU with 64GB memory. The performance is evaluated in terms of the memory consumption and the online search efficiency. Fig. 5 (a) compares the in-memory index as the data size grows, where the x-axis denotes the number of videos in the log scale, and the y-axis measures the index in GB. As we see, the baseline method fails when the data reaches 5 million due to lack of memory. In contrast, our method is scalable and only needs 550MB memory to search 100 million videos. The size of the total inverted index on disk is only 20GB. Fig. 5 (b) compares the online search speed. We create 5 queries, run each query 100 times, and report the mean runtime in milliseconds. A similar pattern can be observed in Fig. 5 that our method is much more efficient than the baseline method and only costs 191ms to process a query on a single core. The above results verify scalability and efficiency of the proposed method.



(a) Index (in GB)

(b) Search Time (in ms)

Figure 5: The scalability and efficiency test on 100 million videos. Baseline method fails when the data reaches 5 million due to the lack of memory. Our method is scalable to 100 million videos.

As a demonstration, we use our system to find relevant videos for commercials. The search is on 800 thousand Internet videos. We download 30 commercials from the Internet, and manually create 30 semantic queries only using semantic visual concepts. See detailed results in the supplementary materials. The ads can be organized in 5 categories. As we see, the performance is much higher than the performance on the MED dataset in Table 2. The improvement is a result of the increased data volumes. Fig. 6 plots the top 5 retrieved videos are semantically relevant to the products in the ads. The results suggest that our method may be useful in enhancing the relevance of in-video ads.

8. CONCLUSIONS

This paper is the first work that addresses a long-lasting challenge of content-based search in 100 million Internet videos. We proposed a scalable solution for large-scale semantic search in video. Our method extends the current capability of semantic video search by a few orders of magnitude while maintaining state-of-the-art retrieval performance. A key in our solution is a novel step called concept adjustment that aims at representing a video by a few salient and consistent concepts which can be efficiently indexed by the modified inverted index. We introduced a novel adjustment model that is based on a concise optimization framework with solid interpretations. We also discussed a solution that leverages the text-based inverted index for video retrieval. Experimental results validated the efficacy and the efficiency of the proposed method on several datasets. Specifically, the experimental results on the challenging TRECVID MED benchmarks validate the proposed method is of state-of-the-art accuracy. The results on the largest multimedia set YFCC100M set verify the scalability and efficiency over a large collection of 100 million Internet videos.

Table 6: Average performance for 30 commercials on the YFCC100M set.

Category	#Ads	Evaluation Metric		
		P@20	MRR	MAP@20
Sports	7	0.88	1.00	0.94
Auto	2	0.85	1.00	0.95
Grocery	8	0.84	0.93	0.88
Traveling	3	0.96	1.00	0.96
Miscellaneous	10	0.65	0.85	0.74
Average	30	0.81	0.93	0.86

Acknowledgments

This work was partially supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. 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).

9. REFERENCES

[1] E. Apostolidis, V. Mezaris, M. Sahuguet, B. Huet, B. Červenková, D. Stein, S. Eickeler, J. L. Redondo Garcia, R. Troncy, and L. Pikora. Automatic fine-grained hyperlinking of videos within a closed collection using scene segmentation. In *MM*, 2014.

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

[3] E. F. Can and R. Manmatha. Modeling concept dependencies for event detection. In *ICMR*, 2014.

[4] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *ECCV*, 2014.

[5] 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.

[6] M. L. Fisher. The lagrangian relaxation method for solving integer programming problems. *Management science*, 50(12):1861–1871, 2004.

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

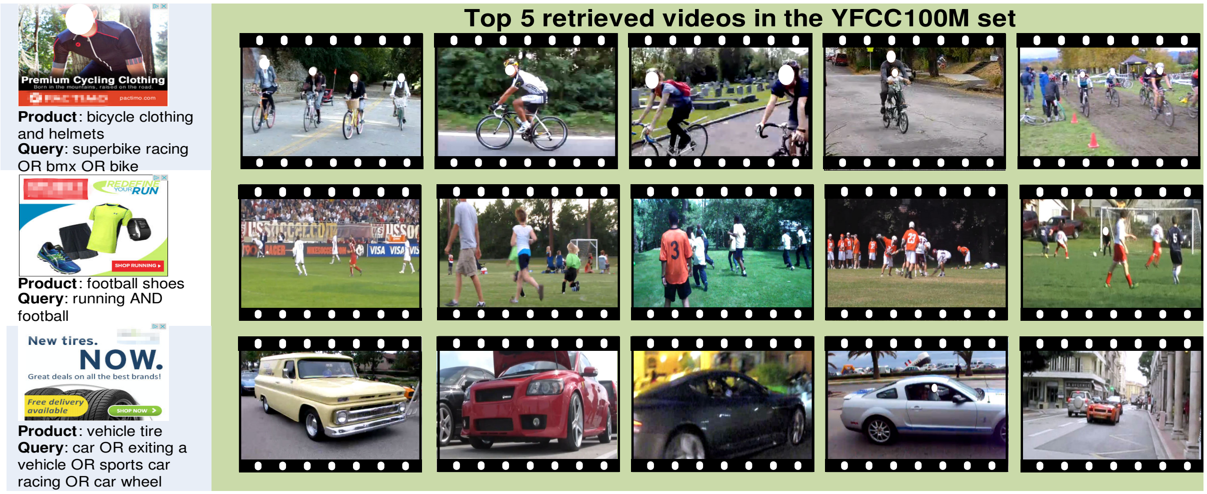

Top 5 retrieved videos in the YFCC100M set
Premium Cycling Clothing
Product: bicycle clothing and helmets
Query: superbike racing OR bmx OR bike
REDEFINE YOUR RUN
Product: football shoes
Query: running AND football
New tires. NOW.
Great deals on all the best brands!
Free delivery available
SHOP NOW
Product: vehicle tire
Query: car OR exiting a vehicle OR sports car racing OR car wheel

Figure 6: Top 5 retrieved results for 3 example ads on the YFCC100M dataset.

[8] M. Grant, S. Boyd, and Y. Ye. CVX: Matlab software for disciplined convex programming, 2008.

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

[10] A. Habibian, T. Mensink, and C. G. Snoek. Videostory: A new multimedia embedding for few-example recognition and translation of events. In *MM*, 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] E. Hatcher and O. Gospodnetic. Lucene in action. In *Manning Publications*, 2004.

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

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

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

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

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

[18] L. Jiang, S.-I. Yu, D. Meng, T. Mitamura, and A. G. Hauptmann. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *ICMR*, 2015.

[19] 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.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

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

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

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

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

[25] D. Moise, D. Shestakov, G. Gudmundsson, and L. Amsaleg. Indexing and searching 100m images with map-reduce. In *ICMR*, 2013.

[26] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *MultiMedia, IEEE*, 13(3):86–91, 2006.

[27] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *MM*, 2004.

[28] 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.

[29] 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.

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

[31] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *arXiv preprint arXiv:1409.0575*, 2014.

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

[33] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

[34] J. Sivic and A. Zisserman. Video google: Efficient visual search of videos. In *Toward Category-Level Object Recognition*, 2006.

[35] J. R. Smith. Riding the multimedia big data wave. In *SIGIR*, 2013.

[36] C. Snoek, K. van de Sande, D. Fontijne, A. Habibian, M. Jain, S. Kordumova, Z. Li, M. Mazloom, S. Pintea, R. Tao, et al. Mediamill at trecvid 2013: Searching concepts, objects, instances and events in video. In *TRECVID*, 2013.

[37] 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.

[38] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

[39] 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.

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

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

[42] 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.

[43] X. Wu, A. G. Hauptmann, and C.-W. Ngo. Practical elimination of near-duplicates from web video search. In *MM*, 2007.

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

[45] S.-I. Yu, L. Jiang, Z. Xu, et al. Informedia@ trecvid 2014 med and mer. In *TRECVID*, 2014.

[46] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.