Learning a Multi-Concept Video Retrieval Model with Multiple Latent Variables

AMIR MAZAHERI, BOQING GONG, and MUBARAK SHAH, Center for Research in Computer Vision, University of Central Florida

Effective and efficient video retrieval has become a pressing need in the "big video" era. The objective of this work is to provide a principled model for computing the ranking scores of a video in response to one or more concepts, where the concepts could be directly supplied by users or inferred by the system from the user queries. Indeed, how to deal with multi-concept queries has become a central component in modern video retrieval systems that accept text queries. However, it has been long overlooked and simply implemented by weighted averaging of the corresponding concept detectors' scores. Our approach, which can be considered as a latent ranking SVM, integrates the advantages of various recent works in text and image retrieval, such as choosing ranking over structured prediction, modeling inter-dependencies between querying concepts, and so on. Videos consist of shots, and we use latent variables to account for the mutually complementary cues within and across shots. Concept labels of shots are scarce and noisy. We introduce a simple and effective technique to make our model robust to outliers. Our approach gives superior performance when it is tested on not only the queries seen at training but also novel queries, some of which consist of more concepts than the queries used for training.

CCS Concepts: • **Computing methodologies** → **Ranking**;

Additional Key Words and Phrases: Video retrieval, multi-concept retrieval, video indexing, structural learning

ACM Reference format:
Amir Mazaheri, Boqing Gong, and Mubarak Shah. 2018. Learning a Multi-Concept Video Retrieval Model with Multiple Latent Variables. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 2, Article 46 (April 2018), 21 pages.
https://doi.org/10.1145/3176647

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center Contract No. D11PC20066. 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 is also supported in part by the NSF Awards No. $IIS 1566511 and No. 1741431.
Authors' addresses: A. Mazaheri and M. Shah, Center for Research in Computer Vision, University of Central Florida, 4328 Scorpius St. Suite 245, Orlando, FL 32816-2365; emails: amirmazaheri@knights.ucf.edu, shah@crcv.ucf.edu; B. Gong, Tencent AI Lab, Bellevue WA 98004; email: boqinggo@outlook.com.

https://doi.org/10.1145/3176647

1 INTRODUCTION

Video data is explosively growing due to the ubiquitous video acquisition devices. Recent years witness a surge of huge video volumes from surveillance, health care, and personal mobile phones, to name a few. From the perspective of the IP traffic, Cisco's white paper on Visual Networking Index reveals several astonishing numbers about the Internet videos (Index 2013, 2014):

- The sum of all forms of videos will be in the range of 80% to 90% of global consumer traffic by 2019.
- It would take an individual more than 5 million years to watch the amount of video that will cross global IP networks each month in 2019.
- Mobile video traffic exceeded 50% of total mobile data traffic in 2012. Note that the global mobile data traffic reached 2.5 exabytes (2.5×10^{18} bytes) per month at the end of 2014.

Indeed, there have been way more videos being generated than what people can watch. Some people have started making fun of this fact; the PetitTube (*www.petittube.com*) randomly plays YouTube (*www.youtube.com*) videos with zero views to its visitors, such that a visitor would become the first one to watch the randomly displayed video.

However, it is not funny at all to see the huge volumes of unwatched videos, but rather alarming, for example, in the public security domain, and rather challenging for video providers to deliver the right videos upon consumers' requests.

By all means, effective and efficient video retrieval has become a pressing need in the era of "big video," whereas it has been an active research area for decades. Following the earlier research on *content*-based video retrieval (Aslandogan and Yu 1999; Hu et al. 2011), the most recent efforts have been mainly spent on (multi-)concept-based video retrieval (Snoek and Worring 2008), an arguably more promising paradigm to bridge the semantic gap between the visual appearance in videos and the high-level interpretations humans perceive from the videos. Multi-Concept-based video retrieval employs a learning algorithm that learns to rank videos in response to the queries consisting of more than one concept. It means that, in both training and testing stages, it handles multiple concepts per query. However, the corresponding learning process may become much harder than content-based or single concept-based video retrieval, since the number of positive examples for the queries degenerates, the number of parameters of the model grows, and uncertainties increase, since different concepts can take place in various parts of a video. These challenges lead us to propose a solution to train a structured model that can be trained and tested with multiple concepts and can deal with the latent inter/intra-shot correlations of the concepts of videos. We refer the readers to the survey articles (Hu et al. 2011; Snoek and Worring 2008) and the annual TRECVID workshops (Over et al. 2014) for a more comprehensive understanding.

A concept corresponds to one or more words or a short description that is understandable by humans. To be useful in automatic video retrieval systems, the concepts (e.g., furniture, beach, etc.) have also to be automatically detectable, usually by some statistical machine-learning algorithms, employing the low-level visual cues (color, texture, etc.) in the videos. Some studies have shown that a rich family of concepts coupled with even poor detection results (10% mean average precision) is able to provide high accuracy results on news video retrieval—comparable to text retrieval on the Web (Hauptmann et al. 2007). Both the richness of the concept set and the performance of the concept detectors are essential. Correspondingly, a plethora of works has been devoted to learning concept detectors (Chang et al. 2007; Dehghan et al. 2014; Qi et al. 2007; Snoek et al. 2013; Yang and Shah 2012; Ye et al. 2015).

Common users have been used to text-based queries for retrieving the target instances in their minds. Equipped with a set of concept detectors, a concept-based video retrieval system is able to accept text descriptions as the queries even when the videos have no textual metadata associated,

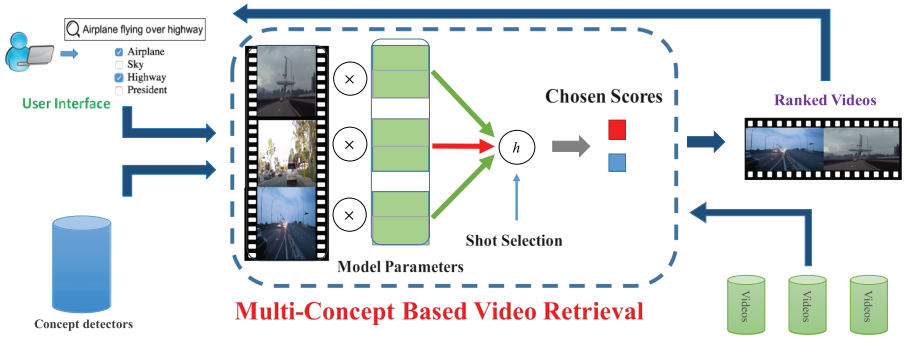

Fig. 1. How to calculate the ranking scores of videos in response to one or more concepts is the central component in many video retrieval systems. It takes as input the multi-concept queries and then returns a ranked list of videos. The multiple concepts in a query could be directly supplied by the users or inferred by the systems from the users' text queries. We apply a set of pre-trained concept detectors on all the shots of the videos. Our model learns to select the most important shots to score a video using multiple latent variables.

thus offering users the same interface for the video retrieval as for document or website retrieval (e.g., Snoek et al. (2007)). As shown in Figure 1, users can directly select concepts from a checklist to compose the queries. Alternatively, the system can also let the user's query be a open-vocabulary text and then translate the queries to a subset of concepts. The latter itself is a very interesting problem to which a full treatment is beyond the scope of this article. Readers who are interested in this problem are referred to Haubold and Natsev (2008), Kennedy et al. (2008), Natsev et al. (2007), Neo et al. (2006), Snoek et al. (2007), Wang et al. (2007), and Wei et al. (2008). In either case, the central operation afterward is to use the resultant subset of concepts to retrieve related videos. In this article, we focus on retrieving whole videos as opposed to segments or shots of videos; however, the developed approach can be conveniently applied to video segments retrieval as well.

Despite being the key component in (multi-)concept-based video retrieval, how to effectively retrieve videos that are related to a subset of concepts is left far from being solved. There is a lack of a principled framework or unified statistical machine-learning model for this purpose. Instead, most existing works take the easy alternative by ranking videos according to the weighted average of the concept detection confidences (Aytar et al. 2008; Mc Donald and Smeaton 2005), where the weights are either uniform or in some systems derived from the similarities between an open-vocabulary user query and the selected concepts. This necessarily fails to take account of the reliabilities of the concept detectors, the relationships between the selected concepts, and the potential contributions from the unselected concepts. Our empirical study actually shows that the unselected concepts can significantly boost the video retrieval performance when they are modeled appropriately.

The objective of this work is to provide a principled model for multi-concept-based video retrieval, where the concepts could be directly provided by the users or automatically selected by the system based on user queries. Figure 1 highlights the main focus of this article in the central panel.

Our approach, which can be considered as a latent ranking SVM (Lan et al. 2012), integrates different advantages of the recent works on text retrieval and multi-attribute-based image retrieval. Particularly, we model the video retrieval as a ranking problem following Grangier and Bengio (2008), as opposed to the structured prediction problem used in Siddiquie et al. (2011) and Yu et al. (2012), to harvest better efficiency and larger modeling capacity to accommodate some latent variables. The latent variables help choose the shots which are the most responsive to the concepts in a user query, without the need of tediously labeling the shots in the training set. We use them

to define the scoring functions both within and across the video shots, closely tracking the unique temporal characteristic of videos. Besides, we incorporate the empirically successful intuitions from multi-attribute-based image retrieval (Siddiquie et al. 2011; Yu et al. 2012) that the inter-dependencies between both selected and unselected concepts/attributes should be jointly modeled. Finally, we introduce a novel 0-1 loss-based early stopping criterion for learning/optimizing our model parameters. This is motivated by the fact that the 0-1 loss is more robust to the outliers than the hinge loss, which is used to formalize the optimization problem.

The proposed model, along with a family of provided scoring functions, accounts for some inevitable caveats of the concept detection results: reliabilities of individual concept detectors, inter-dependencies between concepts, and the correlations between selected and unselected concepts in response to a user query. It expressively improves upon the conventional weighted average of the selected concept detection scores for the multi-concept-based video retrieval. To this end, we stress again that, as the central component in modern video retrieval systems, how to effectively transform the selected concepts to the ranking scores of videos has been long overlooked and is under exploited. More advances and progress on this problem are in need, since they will significantly increase the overall performance of the video retrieval systems.

In the following, we first discuss the related work in Section 2. Section 3.1 presents our video retrieval model given multiple concepts. We then describe how to integrate video-level concept detectors (Section 3.2) and video shot-level detectors (Section 3.3) into the model, followed by experiments in Section 4.

2 RELATED WORK

One of the biggest known challenges in Artificial Intelligence is having systems with understandings close to humans. It has been shown that visual concepts have a promising effect in this understanding of videos or images, especially for retrieval problems (Over et al. 2015; Smeaton 2007; Snoek and Worring 2008). Concepts not only capture the semantics of an environments, they also bridge the gap between low-level visual cues and higher-level representations that humans can understand.

User queries for a retrieval system can include one or more concepts (Campbell et al. 2007; Chang et al. 2005; Haubold and Natsev 2008; Jiang et al. 2015; Kennedy et al. 2008; Natsev et al. 2007; Neo et al. 2006; Snoek et al. 2007; Wang et al. 2007; Wei et al. 2008). The order of the retrieved videos is supposed to be ranked by the relevance of each video to the query concepts. In many works, this is achieved by manually defined similarities and heuristic fusion techniques (Aytar et al. 2008; Li et al. 2007; Mc Donald and Smeaton 2005; Yan and Hauptmann 2003) by mapping the user query concepts to the pool of existing concept detector's scores. Some works like Assari et al. (2014) take into account the positive correlation of concepts occurring in videos. In contrast to existing works, we introduce a principled model to automatically learn both positives and negative correlations of concepts at video and shot-level to rank videos based on queries containing multiple concepts.

Broadly speaking, concept detection encompasses a wide spectrum of topics that are potentially useful for the concept-based video retrieval. Some earlier works (Iyengar and Nock 2003; Smeaton et al. 2009; Snoek et al. 2006) approached this task with tens to hundreds of concepts, while the most recent work has significantly upgraded the scale (Chen et al. 2014; Ye et al. 2015). Concepts can take different forms. For instance, object bank (Li et al. 2010), action bank (Sadanand and Corso 2012), image cell-based detections (Althoff et al. 2012), classemes (Torresani et al. 2010), data-driven concepts (Yang and Shah 2012), sentiment concepts (Chen et al. 2014), events (Ye et al. 2015), and so on. Whereas many concept detectors are trained from manually labeled datasets (Farhadi et al. 2009; Lampert et al. 2014), some other works harvest detectors from the noisy Web data (Li et al. 2013). Our approach benefits from these works. Almost all the existing concept detectors can

be seamlessly integrated into our multi-concept-based video retrieval model by the two types of scoring functions to be described in Section 3.

Work on image retrieval and ranking using multiple attributes (Siddiquie et al. 2011; Yu et al. 2012) is the most relevant to the multi-concept-based video retrieval. However, due to the vast number of video shots in a database, the structural SVM (Tsochantaridis et al. 2004; Yu and Joachims 2009) model used in Siddiquie et al. (2011) and Yu et al. (2012) becomes intractable in our experiments, especially when the annotations are incomplete like the datasets we use in this research. Instead, we develop a simpler ranking model with a variety of scoring functions for our retrieval problem. Our proposed model, which uses multiple latent variables to select shots, is flexible to implement and merges many different scoring functions and also is robust to partially annotated datasets. Also, in contrast to Wu and Worring (2012), which learns to assign a "genre" to each video, our system ranks the videos based on multiple-concepts. Since, a genre can be too specific or too general for what a user is searching for and doesn't match the requirements.

There is a pile of works on learning to rank using the large-margin principle (Grangier and Bengio 2008; Herbrich et al. 1999; Joachims 2002; Lan et al. 2012; Shashua and Levin 2002; Vapnik 2013). The conventional ranking SVMs (Herbrich et al. 1999; Joachims 2002) only learn the ranking function for one query; however, our model is learned for all the queries in the training set, and we show it can be generalized to the queries not seen in training as well without extra training. The model in Grangier and Bengio (2008) is the closest to this work when coupled with our video-level scoring functions, and the latent ranking SVM (Lan et al. 2012) is the closest to ours when coupled with the shot-level scoring functions. However, our scoring functions are particularly tailored to model the characteristics of videos and despite of Lan et al. (2012), we have zero knowledge about the spatial or temporal structure of the concepts in the videos or queries. Besides, we introduce a novel training approach to learn the model parameters by involving the 0-1 loss in an early stopping criterion.

3 APPROACH

In this section, we introduce our main approach to multi-concept-based video retrieval. First, we formalize a ranking model for retrieval problem, and then we describe different scoring functions and how to integrate them into our retrieval model.

3.1 A Ranking Model for Multi-concept Based Video Retrieval

Ranking videos according to one or more concepts selected by the users/systems is the main component in modern video retrieval systems (cf. Figure 1). Whereas most of existing recent works consider a weighted-average of the detection confidences of concepts to rank the videos, we aim to enhance this component by a principled ranking model, which is flexible enough to incorporate different kinds of concept detectors and also can be generalized to unseen queries and deal with temporal dynamics of videos and the correlation of concepts.

We denote all the concepts defined for the system by $\mathcal{Q}$, which users compose queries with and all the videos in the dataset by $\mathcal{V}$. $R(Q) \subset \mathcal{V}$ is the set of all videos that are related to a multi-concept query $Q \subset \mathcal{Q}$. Therefore, an ideal retrieval system must use a learning algorithm to select the best possible subset, named $R(Q)$, among all other subsets of videos from $\mathcal{V}$ such that:

$$\forall S \subset \mathcal{V}, S \neq R(Q),\ R(Q) \text{ is a better output than } S. \tag{1}$$

Directly modeling this notion gives rise to a structured prediction model presented in Siddiquie et al. (2011) and strengthened in Yu et al. (2012). We appreciate that this is perhaps the most natural choice for the retrieval model. There exists $2^{|\mathcal{V}|}$ distinct subsets from $\mathcal{V}$. Unfortunately, this exponential number of subsets makes the computations expensive. Moreover, the expressiveness of the model's scoring function is limited to special forms, to tailor the function to utilize the training

and testing algorithms for structured prediction (Petterson and Caetano 2010; Tsochantaridis et al. 2004). Our own experiments show that it is computationally intractable to use this retrieval model for the shot-level concept detections (cf. Section 3.3).

3.1.1 Retrieval as Ranking. To avoid the high cost of aforementioned structure model, we relax the retrieval problem and solve it as a ranking problem following (Grangier and Bengio 2008). We accommodate multiple latent variables in our model to keep track of the shot-level detections. In particular, the rigorous criterion (Equation (1)) for retrieval is replaced by a less constrained ranking criterion,

$$\forall V_i \in R(Q), \forall V_j \notin R(Q),\ V_i \text{ ranks ahead of } V_j, \tag{2}$$

given V_i and V_j as a pair of relevant and irrelevant videos in the database $\mathcal{V}$ according to the query Q.

Comparing Equation (1) with Equation (2), the former calls for a model to operate over $2^{|\mathcal{V}|}$ subsets of videos, while for the latter we only need a model to assign a ranking score for each video $V \in \mathcal{V}$. The exponential computation complexity makes the solution impractical even for datasets with a few hundreds of videos. We use the following ranking model in this work $\mathcal{F} : \mathcal{Q} \times \mathcal{V} \mapsto \mathbb{R}$,

$$\mathcal{F}(Q, V) = \frac{1}{|Q|} \sum_{q \in Q} f(q, V|\Theta), \tag{3}$$

which breaks down into several ranking scoring functions $f(q, V|\Theta), q \in Q$, each for an individual concept, and Θ denotes the model parameters. We shall write $f(q) \triangleq f(q, V|\Theta)$ in the following for brevity and leave the discussion of the scoring functions to Sections 3.2 and 3.3.

We rank the videos in the database given a multi-concept query Q using $\mathcal{F}$. Top ranked videos in the database will be returned to the user as the video search results. Compared to the retrieval model based on structured prediction (Siddiquie et al. 2011; Yu et al. 2012), our model is not able to optimize the number of videos to output. However, we argue that this does not reduce the usability of our ranking model, considering that common users are used to ranking lists due to text retrieval.

3.1.2 Learning Model Parameters Θ. To train the described model, we follow the ranking SVM (Herbrich et al. 1999; Joachims 2002) strategy. We write our objective function such that

$$\min_{\Theta} \quad \sum_{Q} \frac{1}{|\mathcal{N}(Q)|} \sum_{(i,j) \in \mathcal{N}(Q)} L\left(\mathcal{F}(Q, V_i) - \mathcal{F}(Q, V_j)\right), \tag{4}$$

given $\mathcal{N}(Q)$ as the set of all the pairs of videos V_i and V_j in Equation (2) for the query Q and $L(x) \geq 0$ is a loss function. i and j takes indices of positives and negative videos in the dataset regarding to query Q. One can obtain the pairs from user annotated ranking lists of videos. In Section 4.1, we provide more details about the selection of positive and negatives in our experiments. The loss function will impose some amount of penalty when the ranking scores of a pair of videos violate the ranking constraint of Equation (2).

We exploit two types of loss functions in this work, the hinge loss $L_{\text{hinge}}(x) = \max(1 - x, 0)$ and 0-1 loss $L_{0\text{-}1}(x)$, which takes the value 1 when $x > 0$ and 0 otherwise. *Note that we cannot actually solve the optimization problem with the 0-1 loss by gradient descent;* we instead use it to define a novel early stopping criterion when we solve the problem with hinge loss. Namely, the program stops when the change of the objective function value, computed from the 0-1 loss, is less than a threshold (10^{-10} in our experiments).

As a result, we are able to take advantage of the fact that the 0-1 loss is more robust than the hinge loss when there are outliers in the data. The hinge loss alone would be misled by the outliers and results in solutions that are tuned away from the optimum, while the 0-1 loss helps avoid that

situation by suppressing the penalties incurred by the outliers. Improvements using this novel 0-1 loss stopping criterion, is more sensible in multi-concept video retrieval problem, since for most of the concepts there are many outliers or videos with very different appearances from usual (e.g., cartoons, partial occlusion, distance and point of view, etc.) and also noisy annotations for big datasets. These outliers make the training stage much more harder. Indeed, the novel stopping criterion by the 0-1 loss significantly improves the results of hinge loss in our experiments.

Note that the 0-1 loss-based stopping criterion is another key point clearly differentiating our approach from Grangier and Bengio (2008), which motivates our ranking model. In addition, we introduce a family of new scoring functions for different concept detections, especially the one with latent variables in Section 3.3.

3.2 Video-level Concept Detection

It is a common practice that the concept detection results $\boldsymbol{\phi}(V)$ over each video $V \in \mathcal{V}$ are computed off-line and stored somewhere to speed up the responding to the users' queries. We use $\boldsymbol{\phi}$ as the shorthand of $\boldsymbol{\phi}(V)$. Note that $\boldsymbol{\phi}$ is a $|\mathcal{Q}|$-dimensional vector whose entry ϕ_q corresponds to the detection confidence of the concept q (in a video V). We next describe how to use $\boldsymbol{\phi}$, the video-level concept detection results, to design the scoring functions $f(q), q \in Q \subset \mathcal{Q}$ (cf. Section 3.1). We start from the weighted average approach that prevails in the existing video retrieval works.

3.2.1 Weighted Average. Recall that the overall scoring function $\mathcal{F}(Q, V)$ breaks down into several individual functions $f(q, V|\boldsymbol{\Theta}) \triangleq f(q), q \in Q$, each of which accounts for one concept (Equation (3)). A common practice to rank the videos given a multi-concept query Q is by the average of the corresponding concept detection confidences:

$$f^{\mathrm{V}}_{\mathrm{avg}}(q) = \phi_q \triangleq \langle \mathbf{1}^q, \boldsymbol{\phi} \rangle, \tag{5}$$

where the weights are simply uniform and $q \in Q$. $\mathbf{1}^q \in \{0, 1\}^{|Q|}$ denotes a one-hot vector, which is 1 at the qth entry and 0's else, or, when the query is an open-vocabulary text, the weights could be the similarities inferred between the concepts in Q and the user query (Kennedy et al. 2008; Wang et al. 2007). We include the uniform weights in our experiments without loss of generality.

The mentioned average method fails to model the correlations between the concepts in Q, and also the correlations between Q and the remaining unselected concepts in $\mathcal{Q}$. To make it more clear, we have re-written the score function by the one-hot vectors on the rightmost of Equation (5). Therefore, the model parameters become $\boldsymbol{\Theta} = (\mathbf{1}^1, \mathbf{1}^2, \cdots, \mathbf{1}^{|\mathcal{Q}|})^T = I \in \mathbb{R}^{|\mathcal{Q}| \times |\mathcal{Q}|}$, i.e., actually an identity matrix. The entry $\boldsymbol{\Theta}_{qp}$, which is supposed to encode the relationship of concepts q and p, is 0 in the weighted average (Equation (5)).

3.2.2 Encoding Concept Correlations. To this end, the natural extensions to the weighted average scoring function are the following:

$$f^{\mathrm{V}}_{\text{corr-1}}(q) = \langle \boldsymbol{\theta}^q, \boldsymbol{\phi} \rangle, \text{ such that } \boldsymbol{\theta}^q_p = 0 \text{ if } p \notin Q, \tag{6}$$

and

$$f^{\mathrm{V}}_{\text{corr-2}}(q) = \langle \boldsymbol{\theta}^q, \boldsymbol{\phi} \rangle, \text{ such that } \boldsymbol{\theta}^q \in \mathbb{R}^{|\mathcal{Q}|}, \tag{7}$$

where $f^{\mathrm{V}}_{\text{corr-1}}(q)$ considers the contributions from the other concepts $p \in Q$ when it scores the concept q for a video V. Indeed, the existence of "computer" could affect the confidence of "furniture" when both are selected to the query $Q \subset \mathcal{Q}$. The other function $f^{\mathrm{V}}_{\text{corr-2}}(q)$ further considers all the other concepts not in the query, when it scores for instance the concept $q =$ "furniture" in a video, capturing the case that the lack of "beach" may reinforce the confidence about "furniture."

Note that the parameters $\Theta = (\boldsymbol{\theta}^1, \cdots, \boldsymbol{\theta}^{|\mathcal{Q}|})^T$ become a full matrix now, offering more modeling flexibilities.

3.3 Shot-level Concept Detection

Many concept detectors are designed to take frames or video shots as input (Althoff et al. 2012; Sadanand and Corso 2012; Ye et al. 2015). We split any given video V in the database into H shots. This can be done by applying a shot detector (Paul et al. 2011) or uniformly select the shot boundaries inside a full sequence video. We compute and store the concept detection results $\boldsymbol{\phi}^h \in \mathbb{R}^{|\mathcal{Q}|}, h = 1, \ldots, \mathrm{H}$ for all the concepts $\mathcal{Q}$ over the shots of the video. Shot-level detections provide more detailed information about the videos than video-level concept detections considering the dynamics of the video and sudden changes of the scene. Therefore, we propose our novel family of scoring functions, which leverage the shot-level detection scores.

3.3.1 Scoring the Best Shot for the Querying Concepts. One potential benefit we can have from the shot-level concept detections is that, among all the shots of a video V, we can select the most informative shot for the scoring function:

$$f^{\mathrm{S}}_{\text{latent}}(q) = \max_{h \in \{1,2,\ldots,\mathrm{H}\}} \left\langle \boldsymbol{\theta}^q, \boldsymbol{\phi}^h \right\rangle, \tag{8}$$

where the model parameters $\boldsymbol{\theta}^q \in \mathbb{R}^{|\mathcal{Q}|}$, which correspond to the concept $q \in Q \subset \mathcal{Q}$, represent the contributions to q from all the concepts within the shot, which will be selected by the latent variable $h \in \{1, 2, \ldots, \mathrm{H}\}$.

We argue that this formulation can better model the *negative* correlations between concepts, if any, than the $f^{\mathrm{V}}_{\text{corr-2}}(q)$ defined over the video level. Indeed, consider a set of negatively correlated concepts (e.g., "beach," "furniture," etc.). They could all have very strong responses across a video. For instance, a tourist may capture a video within a hotel room and then shift to the beach outside. As a result, both "beach" and "furniture" will be detected with high confidences in the video but they are exclusive over a single shot. As a result, the video-level scoring function may be confused by this video but $f^{\mathrm{S}}_{\text{latent}}(q)$ scores a single best shot and is thus robust to this scenario.

3.3.2 Scoring Best Shot for Each Querying Concept. While $f^{\mathrm{S}}_{\text{latent}}(q)$ well captures the negative concept correlations by possibly negative off-diagonal entries, it may be inadequate to track the *positive* correlations between different concepts by focusing on only a shot; very few concepts could appear simultaneously in a single shot. We thus compensate it by a second term:

$$f^{\mathrm{VS}}_{\text{latent}}(q) = \max_{h \in \{1,\ldots,\mathrm{H}\}} \left\langle \boldsymbol{\theta}^q, \boldsymbol{\phi}^h \right\rangle + \sum_{p \in Q} \max_{g \in \{1,\ldots,\mathrm{H}\}} v^q_p \phi^g_p, \tag{9}$$

where $\max_g v^q_p \phi^g_p$ max-pools the confidences of each concept across all the shots of video V. Note that we therefore provide two complementary types of modeling capabilities in $f^{\mathrm{VS}}_{\text{latent}}(q)$. The first term is robust to the concepts that are negatively correlated with q and the products in the second term strengthen the detection score of concept q from some positively correlated concepts in the video. The model parameters $\boldsymbol{\theta}^q$ and $\boldsymbol{v}^q$ are learned by solving Equation (4) with sub-gradient descent. Details are given as follows.

3.4 Optimization

We use (Sub-)Gradient Descent(SGD) as a solver to learn our model parameters Θ. As discussed in Section 3.1.2, the loss function L is a non-zero and positive value for each pair $\{(V_i, V_j)\}$, in which the negative video V_j has higher ranking score than the positive V_i. As a result, the gradients for those pairs are non-zero and zero for the others.

Denoting by

$$\mathcal{S}_j = \frac{\partial L}{\partial \mathcal{F}(Q, V_j)} \times \frac{\partial \mathcal{F}(Q, V_j)}{\partial \boldsymbol{\Theta}}, \tag{10}$$

we thus have the overall gradients of Equation (4) by

$$\sum_{Q} \frac{1}{|\mathcal{N}(Q)|} \sum_{(i,j)\in\mathcal{N}(Q)} \left(\mathcal{S}_i - \mathcal{S}_j\right). \tag{11}$$

Note that the model parameters $\boldsymbol{\Theta}$ consist of two parts $(\boldsymbol{\theta}, \boldsymbol{v})$, corresponding to the two terms of $f^{\text{VS}}_{\text{latent}}$ (cf. Equation (9)), respectively. We compute the gradients with respect to the first part θ using the $softmax$ derivation to approximate a smooth gradients, as suggested by Ping et al. (2014):

$$\frac{\partial \mathcal{F}(Q, V_j)}{\partial \boldsymbol{\theta}} = \sum_{h\in\{1,2,\ldots,\text{H}\}} \frac{e^{\langle \boldsymbol{\theta}^q, \boldsymbol{\phi}^h \rangle} \boldsymbol{\phi}^h}{\sum_{j\in\{1,2,\ldots,\text{H}\}} e^{\langle \boldsymbol{\theta}^q, \boldsymbol{\phi}^j \rangle}}. \tag{12}$$

We write out the gradients with respect to the second part, $\boldsymbol{v}$, over different dimensions of $\boldsymbol{v}$. Recall that he second term of $f^{\text{VS}}_{\text{latent}}$ (cf. Equation (9)), $\max_g v^q_p \phi^g_p$, max-pools over all the shots of a video for each single concept. As a result, we have the following:

$$\frac{\partial \mathcal{F}}{v^q_p} = \phi^{g*}_p, \quad p = 1, 2, \ldots, \text{H}, \; q = 1, \ldots, \text{H}, \tag{13}$$

where $g*$ is determined by $g* \leftarrow \max_g v^q_p \phi^g_p$.

4 EXPERIMENTS

This section presents our experiments to evaluate the proposed multi-concept video retrieval model along with the various scoring functions. We first describe the experiment setup and then report the retrieval results. We further give some detailed analyses and qualitative results.

4.1 Experiment Setup

We use two separate datasets in our experiments, respectively, for video retrieval and training the concept detectors. We further exploit four sets of multi-concept queries. Two sets consist of 50 queries each in the form of concept pairs, one for training and testing and the other just for testing. Two other sets contain 50 triplets and 30 single concept queries just for testing, respectively. We train our model using only the first set of queries on the training set and then test it by all four sets of queries on the test set.

4.1.1 The IACC.2.B Dataset for Video Retrieval. We mainly test our approach over the IACC.2.B dataset, which is the test set used in the Semantic Indexing (SIN) task of TRECVID 2014 (Over et al. 2014) challenge. The dataset comprises 2,371 Internet videos that are "characterized by a high degree of diversity in creator, content, style, production qualities, original collection device/encoding, language, and so on" (Over et al. 2014). The video durations range from 10 seconds to 6.4min with the mean of 5min. Standard shot partitions (106,000 shots in total) are provided by the dataset and 30 concepts are annotated for the shots. We use TRECVID 2015 SIN Task's test set, named IACC.2.C dataset, as the second dataset for more experiments as a correctness proof of our approach. All the settings, such as splits ratios, are the same as IACC.2.B.

We randomly split IACC.2.B to 712 videos as the training set, 474 videos as the validation set and 1,185 videos as the test set. We select 50 pairs of concepts from the total $\binom{30}{2}$ possible pairs. We select them based on the number of positive examples in the training set and call them the

Table 1. The Multi-concept Queries Used in Our Experiments

	Queries
50 Concept Pairs	(Chair, Computers), (Boat/Ship, Oceans), (Classroom, Computers), (Instrumental_Musician, Singing), (Chair, Classroom) (Boat/Ship, Running), (Chair, Nighttime), (Boat/Ship, Quadruped), (Bicycling, Forest), (Chair, Hand), (Chair, Flags) (Boat/Ship, Forest), (Nighttime, Singing), (Cheering, Singing), (Forest, Lakes), (Chair, Telephones), (Running, Stadium) (Chair, Forest), (Beach, Boat/Ship), (Oceans, Quadruped), (Forest, Quadruped), (Beach, Quadruped), (Running, Forest) (Bridges, Forest), (Boat/Ship, Bridges), (Instrumental_Musician, Nighttime), (Highway, Nighttime), (Beach, Oceans) (Bus, Highway), (Bus, Chair), (Nighttime, Forest), (Highway, Forest), (Computers, Telephones), (Nighttime, Running) (Bridges, Highway), (Cheering, Flags), (Cheering, Instrumental_Musician), (Cheering, Nighttime), (Forest, Oceans) (Chair, Highway), (Chair, Quadruped), (Boat/Ship, Lakes), (Running, Quadruped), (Nighttime, Flags), (Bridges, Chair) (Boat/Ship, Nighttime), (Demonstration_Or_Protest, Flags), (Airplane, Boat/Ship), (Boat/Ship, Chair), (Chair, Running)
50 Concept Triplets	(Ocean, Quadruped, Boat/ship), (Chair, Classroom, Computers), (Beach, Boat/Ship, Quadruped), (Cheering, Boat/Ship, Flags) (Chair, Computers, Telephones), (Cheering, Instrumental_Musician, Singing), (Instrumental_Musician, Nighttime, Singing) (Forest, Boat/Ship, Oceans), (Lakes, Boat/Ship, Oceans), (Beach, Ocean, Boat/Ship), (Quadruped, Beach, Ocean) (Forest, Lakes, Quadruped), (Boat/Ship, Bridges, Forest), (Chair, Nighttime, Forest), (Bridges, Forest, Lakes) (Boat/Ship, Forest, Quadruped), (Boat/Ship, Forest, Lakes), (Running, Oceans, Boat/Ship), (Highway, Boat/Ship, Bridges) (Running, Beach, Oceans), (Quadruped, Running, Forest), (Cheering, Instrumental_Musician, Nighttime) (Boat/Ship, Nighttime, Forest), (Chair, Ocean, Boat/Ship), (Running, Oceans, Quadruped), (Chair, Highway, Bridges) (Beach, Forest, Oceans), (Bridges, Oceans, Boat/Ship), (Lakes, Bridges, Boat/Ship), (Running, Beach, Boat/Ship) (Cheering, Singing, Flags), (Demonstration, Bus, Flags), (Lakes, Boat/Ship, Quadruped), (Bus, Chair, Computers) (Boat/Ship, Flags, Oceans), (Forest, Bus, Boat/Ship), (Quadruped, Forest, Bicycling), (Oceans, Forest, Lakes) (Bridges, Chair, Boat/Ship), (Boat/Ship, Bridges, Bus), (Forest, Boat/Ship, Highway), (Cheering, Nighttime, Singing) (Quadruped, Ocean, Forest), (Flags, Demonstration, Nighttime), (Bus, Bridge, Highway), (Boat/Ship, Nighttime, Oceans) (Chair, Boat/Ship, Forest), (Forest, Chair, Quadruped), (Highway, Bus, Chair), (Bridges, Highway, Forest)

The concept-pair queries shown in this table are mentioned as **Seen Queries** and have been used for training the system.

seen queries, which always will be used in training and in one experiment as the test. A video is a positive or related sample for a query when there is at least one shot annotated as positive for all the concepts in that query. This results in minimally 27, maximally 86, and on average 44 out of the 1,185 videos in the database (i.e., the test set) related to the concept-pair queries. See Table 1.

Additionally, we build a set of concept-triplet queries with the size of 50. On average each concept-triplet query has 24 related videos. As a matter of fact, by increasing the number of concepts, the retrieval task becomes more challenging due to less number of positive examples, and we show in our experiments that our method can handle them very well. We never use concept-triplets as training and use them just for a test as unseen queries. In Table 1, we show our set of concept-triplets.

4.1.2 Evaluation. We use one of the most popular metrics in information retrieval, Normalized Discounted Cumulative Gain (NDCG) (Järvelin et al. 2008), to evaluate the ranking lists returned by our model in response to the multi-concept queries. Given a ranking list for query Q, NDCG is calculated by

$$\text{NDCG@}k = \frac{1}{Z} \sum_{j=1}^{k} \frac{G[j]}{1 + \log_2 j}, \tag{14}$$

given the gain function $G[j] = \text{rel}(j)^2$ and j as the rank of a video that has $\text{rel}(j)$ number of concepts that video shares with the query Q. By changing the power in the gain function, we can tune how important is for us that a retrieved video contains all the desired concepts. The partition Z is the ideal gain value computed by ground truth, which will produce an ideal ranking list. As a result, any NDCG@k value is normalized between 0 and 1. We shall report the results at $k = 5, 10, \ldots, 50$ in the following experiments.

4.1.3 Concept Detectors. Learning robust concept detectors has a rich literature (Qi et al. 2007; Snoek et al. 2013; Ye et al. 2015). All kinds of concept detectors can be potentially employed in our retrieval model. We train our independent concept detectors following the practice of Mazaheri et al. (2015).

In particular, we train 60 independent detectors from the training data (IACC.1.tv10.training, IACC.1.A, IACC.1.B, and IACC.1.C) of the TRECVID 2014 SIN task (Over et al. 2014) for the concepts with key frame annotations, including the 30 concepts annotated in IACC.2.B. To this end, we extract dense SIFT (Lowe 2004) (DSIFT) and Convolutional Neural Network (CNN) features (Krizhevsky et al. 2012) from the annotated (both positive and negative) key frames for the 60 concepts. The DSIFT features computed using VLFeat (Vedaldi and Fulkerson 2010) toolbox are encoded by Fisher vectors (Perronnin et al. 2010) as an image representation and then input to linear SVMs. For CNN features, we use the activations of "relu6" and "fc7" layers as two types of image representations, train SVMs with histogram intersection kernels for each of them, and then average the detection scores of the two types of SVMs. Overall, we thus harvest two complementary detection confidences for any concept, one from the DSIFT and the other from the CNN features. They are both transformed to probabilities using the Platt calibration. At the testing stage, we first average them to obtain the concept detection confidences for each frame, then max-pool the scores within a shot to have the shot-level results $\boldsymbol{\phi}^h$, and finally arrive at the video-level concept detection results $\boldsymbol{\phi}$ by max-pooling.

4.1.4 Practical Considerations in Implementation. To prevent the model from overfitting, we have employed an $L2-norm$ as regularizer. In particular, we add $\sum_{q\in Q} \lambda \left\|\boldsymbol{\theta}^q\right\|_2^2 + \gamma \left\|\boldsymbol{v}^q\right\|_2^2$ term to regularize the optimization problem in Equation (4). We have used the validation set to tune the hyper-parameters λ and γ. Note that $\gamma = 0$ except for the scoring function $f^{\text{VS}}_{\text{latent}}(q)$. We also remove all the videos that have no shot annotated as negative or positive for the given query.

4.2 Comparison Results

We compare different scoring function results in Table 2. For this experiment, we have used IACC.2.B dataset and the 50 pair-concepts queries shown in Table 1 for both training and testing. We evaluate using NDCG@k where $k = 5, 10, \ldots, 50$ for different columns. The left-hand column shows the mean of NDCG for all k values.

Following the common practice in the existing concept-based video retrieval systems, we empirically test a variety of fusion methods (Campbell et al. 2007; Ishikawa et al. 2013; Yan and Hauptmann 2003) as the (old) baselines-our approach offers a new set of simple yet more advanced ranking scheme for the multi-concept-based video retrieval. Probably because our detectors output probabilities after the Platt calibration, the average operation in $f^{\text{V}}_{\text{avg}}$ performs the best among the fusion techniques discussed in Campbell et al. (2007). We thus only show the results of $f^{\text{V}}_{\text{avg}}$ and the second best, PicSOM (Ishikawa et al. 2013), in the rows tagged by "Common practice" and "PicSOM 2013", respectively, in Table 2. The PicSOM fusion strategy (Ishikawa et al. 2013) involves a convex combination of the product and the average of the querying concepts' detection scores. Also, we used an another common technique as explained in Assari et al. (2014) to capture just positive correlation between pairs of concepts, using a Co-occurrence matrix of them built in training stage.

We further include ranking SVM (Joachims 2002) and TagProp (Guillaumin et al. 2009) in the table as more advanced baselines. Both take as input the video-level representations; they are not able to handle the set of shot-level features in each video. We use two types of inputs, the video-level concept detection scores, and CNN features as the video representations to train the TagProp and ranking SVM models. Note that we train a ranking SVM model for each of the pair-concept queries. TagProp is a state-of-the-art image tagging algorithm. It uses K-nearest neighbor and

Table 2. Comparison Results of Different Scoring Functions in Pair-concept-based Video Retrieval

Baselines

Functions	NDCG@5	@10	@15	@20	@25	@30	@35	@40	@45	@50	Mean
PAMIR (Grangier and Bengio 2008)	04.3%	04.2%	04.1%	04.3%	04.7%	05.2%	05.7%	06.1%	06.5%	06.7%	05.22%
Fast0Tag (Zhang et al. 2016)	07.6%	07.2%	07.3%	07.7%	08.0%	08.2%	08.6%	08.9%	09.0%	09.1%	08.2%
Common practice $f^{\mathrm{V}}_{\mathrm{avg}}$	62.6%	57.1%	55.6%	56.1%	57.5%	58.8%	59.7%	61.0%	62.0%	62.6%	59.3%
TagProp (Guillaumin et al. 2009)	30.0%	27.3%	25.6%	26.8%	27.7%	28.6%	29.4%	30.1%	30.8%	31.4%	28.8%
Rank-SVM (Chapelle and Keerthi 2010)	57.9%	52.9%	52.2%	52.6%	54.3%	55.4%	56.5%	56.8%	57.7%	58.1%	55.5%
Co-occurrence (Assari et al. 2014)	59.4%	50.7%	48.6%	49.5%	51.8%	53.4%	54.9%	55.6%	56.4%	57.4%	53.8%
PicSOM 2013 (Ishikawa et al. 2013)	63.0%	57.1%	55.5%	55.9%	57.3%	58.1%	59.2%	60.5%	61.5%	62.1%	59.0%
$\|Q\|$= **30 concepts**											
Video-level $f^{\mathrm{V}}_{\mathrm{corr\text{-}1}}$	61.6%	57.0%	55.5%	55.8%	57.0%	58.1%	59.5%	60.5%	61.5%	62.2%	58.9%
Video-level $f^{\mathrm{V}}_{\mathrm{corr\text{-}2}}$	64.8%	59.2%	57.4%	57.6%	58.9%	60.3%	61.5%	62.3%	63.1%	63.3%	60.9%
Shot-level $f^{\mathrm{S}}_{\mathrm{latent}}$	*68.2%*	*61.0%*	*59.5%*	59.5%	61.2%	62.1%	63.6%	64.9%	65.8%	66.1%	*63.2%*
Shot-Video-level $f^{\mathrm{VS}}_{\mathrm{latent}}$	62.9%	58.8%	58.3%	*60.0%*	*61.8%*	*63.1%*	*64.7%*	*65.4%*	*66.2%*	*67.1%*	62.8%
$\|Q\|$= **60 concepts**											
Video-level $f^{\mathrm{V}}_{\mathrm{corr\text{-}1}}$	61.6%	57.0%	55.5%	55.8%	57.0%	58.1%	59.5%	60.5%	61.5%	62.2%	58.9%
Video-level $f^{\mathrm{V}}_{\mathrm{corr\text{-}2}}$	64.8%	58.8%	57.3%	57.3%	58.8%	60.0%	61.1%	62.4%	62.9%	63.2%	60.7%
Video-level with RBF Distance $f^{\mathrm{V}}_{\mathrm{corr\text{-}2}}$	64.0%	59.3%	57.3%	57.3%	58.7%	60.1%	61.1%	62.4%	62.9%	63.3%	60.6%
Shot-level $f^{\mathrm{S}}_{\mathrm{latent}}$	67.6%	61.8%	59.8%	59.7%	61.5%	62.6%	64.1%	65.1%	65.8%	66.5%	63.5%
Shot-Video-level $f^{\mathrm{VS}}_{\mathrm{latent}}$	69.8%	**63.8%**	**61.7%**	**60.9%**	**63.0%**	**64.1%**	**65.4%**	**66.4%**	**66.8%**	**67.4%**	**64.9%**
Shot-Video-level with RBF Distance $f^{\mathrm{VS}}_{\mathrm{latent}}$	**70.2%**	62.2%	60.0%	60.3%	62.1%	63.2%	64.5%	65.4%	66.1%	67.1%	64.1%

metric learning to propagate the tags of training examples to any testing instance. We report the best results for each after parameter tuning on our validation set. Probably because our training set is relatively small, the TagProp results are very low. Ranking SVM gives comparable results with the other fusion techniques.

Moreover, we adopt Fast0Tag (Zhang et al. 2016) and Passive Aggressive Model for Image Retrieval (PAMIR) (Grangier and Bengio 2008) methods on all three datasets (cf. Table 4). Both of these methods try to find a mapping between image feature and concepts. Fast0tag uses the popular pre-trained word2vec (Pennington et al. 2014) model that is trained on a large text corpus, to embed each concept in a high dimensional space where word (concept) similarities can be measured more accurately. It maps each image to its corresponding unique direction to rank concept scores. PAMIR also follows the same idea; however, this method learns to represent each query of multi-concept by a single vectors.

There are four types of scoring functions in our approach, $f^{\mathrm{V}}_{\mathrm{corr\text{-}1}}$ and $f^{\mathrm{V}}_{\mathrm{corr\text{-}2}}$ accounting for the video-level concept detections and $f^{\mathrm{S}}_{\mathrm{latent}}$ and $f^{\mathrm{VS}}_{\mathrm{latent}}$ for the shot-level concept detections. We learn these models' parameters by the 0-1 loss-based early stopping and the 50 pair-concept queries (cf. Table 1) using the videos in the training set. Experiments comparing the hinge loss and the 0-1 loss are presented in Section 4.3.

To further study our model's behavior, we replace the linear distance between positive and negative neighbors described in Equation (4) with a non-linear distance. To be more specific, we replace $\mathcal{F}(Q, V_i) - \mathcal{F}(Q, V_j)$ in Equation (4), with $sign(\mathcal{F}(Q, V_i) - \mathcal{F}(Q, V_j)).(1 - K(\mathcal{F}(Q, V_i), \mathcal{F}(Q, V_j)))$, where K is the Radial Basis Function (RBF) similarity function. We keep the *sign* of the distance, since we are solving a ranking problem and the order of the ranked samples must be preserved.

Comparison. There are $|\mathcal{Q}| = 30$ concepts labeled for our video database $\mathcal{V}$, which are drawn from the IACC.2.B dataset. All our queries are constructed from these concepts such that we have the ground-truth ranking list for evaluation. We first show the video retrieval results in the top half of Table 2. The variations of our model with different scoring functions all improve the common practice $f^{\mathrm{V}}_{\mathrm{avg}}$. The margins between $f^{\mathrm{V}}_{\mathrm{avg}}$ and our latent shot-level functions are especially significant.

The Benefit of More Concepts. Though the queries are built from the vocabulary of 30 concepts, we are actually able to harvest more concept detectors from another independent dataset, TRECVID 2014 SIN task training set. Our model is flexible to include them by expanding the concept detection vectors $\boldsymbol{\phi}$ (see Section 3). The bottom half of Table 2 shows the results corresponding to 60 concept detectors. We see that the observations about the relative performances of the model variations from the $|\mathcal{Q}| = 30$ concepts still hold. In addition, the video retrieval results using the shot-level scoring functions have been significantly improved over those of the 30 concepts. This is in accordance with our intuition as well as the results in Yu et al. (2012). Indeed, the inter-dependences of more concepts may provide better information for our scoring functions and make them more robust to the unreliable concept detection confidences.

Note that, however, introducing more concepts does not change the results of the video-level scoring function $f^{\mathrm{V}}_{\text{corr-2}}$ too much, $f^{\mathrm{V}}_{\text{corr-1}}$ is not affected by extra concepts. We argue that this is mainly due to the fact that our detectors are not developed for the video-level concept detections. For the future work, it will be interesting to see whether more video-level concept detectors, such as the action classifiers (Hou et al. 2014; Wang and Schmid 2013), can benefit our video-level function $f^{\mathrm{V}}_{\text{corr-2}}$ as well. Another interesting direction would be to pursue the weak attributes/concepts in the video retrieval task (Yu et al. 2012).

4.3 The Effect of the 0-1 Loss

We study the effect of the novel 0-1 loss-based stopping criterion in this section. Figure 2 shows the retrieval results of both the video-level scoring function $f^{\mathbf{V}}_{\text{corr-2}}$ and shot-level function $f^{\mathrm{VS}}_{\text{latent}}$, respectively, with and without using the 0-1 loss in the optimization. We can see that coupling the 0-1 loss as a stopping criterion with the hinge loss in optimization significantly improves the performance of the hinge loss alone for both types of scoring functions. This is not surprising. The 0-1 loss is advantageous over the hinge loss, especially when there are "difficult" positive-negative pairs that heavily violate the ranking constraint in the training/optimization stage. The hinge loss would penalize more those pairs and consequently ignore the other pairs, but the 0-1 loss is resilient to those pairs. Although in practice we cannot fully harvest the appealing modeling power of the 0-1 loss due to the gradient descent, our results in Figure 2 verify that the nice properties of the 0-1 loss can be transferred indirectly by defining the new stopping criterion with the 0-1 loss. And also, it shows a better performance for top retrieved videos (smaller k values in Figure 2), where the 0-1 stopping criterion shows a lot of improvement. We also provide qualitative results in Figure 6 for more clarifications.

4.4 Generalizing Out to Unseen Queries

We expect our model to generalize well to other multi-concept queries. To demonstrate this, we train our system with pair-concept queries shown in Table 1 and test it on three other sets of queries of which none of them have any query in common with the ones used in training: (a) 30 single-concept queries, (b) 50 new pair-concept queries, and (c) 50 new triple-concept queries. None of them are used to train our model. The 50 concept triplets are shown in Table 1. Figure 3 shows the retrieval results using different variations of our model. We can see our model with the shot-level scoring functions $f^{\mathrm{S}}_{\text{latent}}$ and $f^{\mathrm{VS}}_{\text{latent}}$ performs quite well upon the new queries.

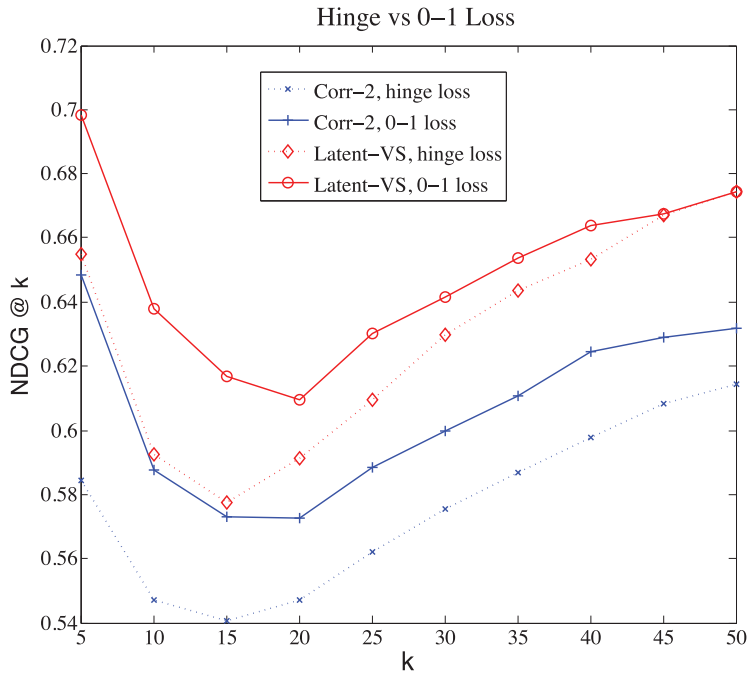

Fig. 2. The effects of the 0-1 loss-based stopping criterion in optimization. For both the video-level scoring function $f^{\text{V}}_{\text{corr-2}}$ and shot-level function $f^{\text{VS}}_{\text{latent}}$, the introduction of the 0-1 loss significantly improves the performance of the hinge loss.

Table 3. Results for TRECVID 2014 SIN Challenge

Method	$f^{\text{V}}_{\text{avg}}$	$f^{\text{VS}}_{\text{latent}}$
IACC.2.B (Full) - TRECVID SIN 2014	24.01	—
IACC.2.B (Half) - TRECVID SIN 2014	24.56	24.80

The results are not only significantly better than the simple average, but also comparable to those for the previously seen pair-concept queries (cf. Table 2). The video-level scoring function $f^{\text{V}}_{\text{corr-2}}$ unfortunately degrades and gives similar or worse performance compared to the averaging baseline $f^{\text{V}}_{\text{avg}}$. It implies that our multi-latent variable-based scoring function is able to generalize the trained model and use it to retrieve unseen queries.

In a complementary experiment, we show our model can be used on TRECVID SIN challenge. Even though this challenge is based on short video shots and single concept queries retrieval, an improvement is expected due correlation of concepts that our model captures. For this experiment, we use the same testing pipeline. Shots are considered as videos and each frame as a single shot. The results are given in Table 3. We use the same settings as explained in Section 4.1 and the learned model is the same as used in other experiments. To give an insight into performance of the independent concept detectors, we try them on the full set as well. The numbers are reported using mean InfAP (Yilmaz and Aslam 2006) over 30 concepts. The first 2,000 retrieved shots for each concept are considered. Note that our model is not applicable to full dataset due to 50% usage of that in training stage.

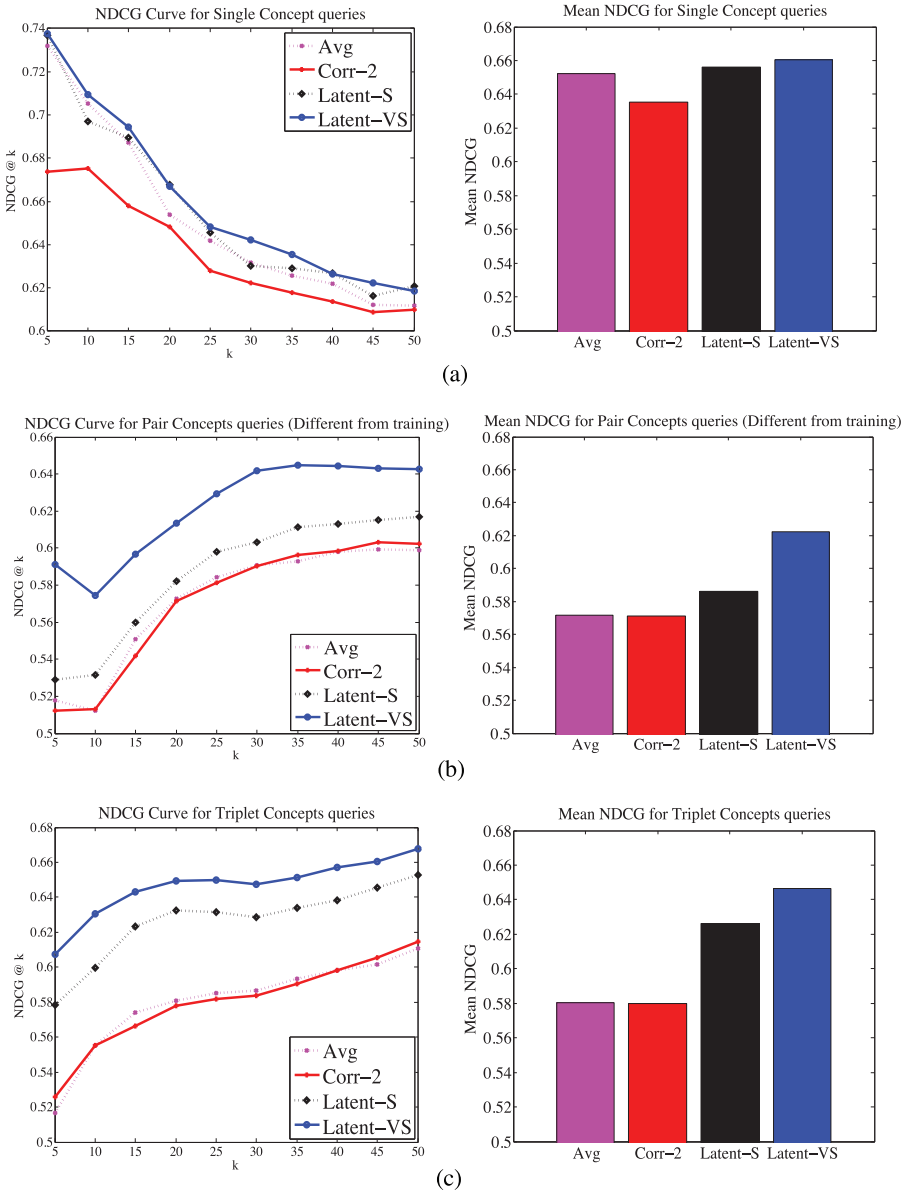

Fig. 3. *Previously unseen* queries: (a) single-concept queries, (b) pair-concept queries, and (c) triple-concept queries.

4.5 Extensive Experiments on IACC.2.C Dataset

To have an extensive set of experiments, we use IACC.2.C dataset. As explained in Section 4.1.1, we use similar setting for extracting queries and also training/validation/test sets as IACC.2.B. We go further integrate IACC.2.C and IACC.2.B to build one super dataset. In Table 4, we show a side by side comparison between three different datasets as well as different scoring functions and baselines. Clearly, our multi-latent variables method gives the superior performance in all three cases.

Table 4. Baseline Averages of NDCG@5-50 on Three Different Datasets

Dataset	IACC.2.B	IACC.2.C	IACC.2.B + C
PAMIR	5.2%	38.3%	18.0%
Fast0Tag	8.2%	30.6%	18.4%
TagProp	28.8%	24.1%	12.3%
Rank-SVM	55.5%	26.4%	39.5%
Co-occurrence	53.8%	44.4%	30.2%
PicSOM	59.0%	53.2%	40.3%
$f^{\mathrm{V}}_{\mathrm{avg}}$	59.3%	53.7%	40.4%
$f^{\mathrm{V}}_{\mathrm{corr\text{-}2}}$	60.7%	50.1%	37.8%
$f^{\mathrm{VS}}_{\mathrm{latent}}$	**64.9%**	**56.9%**	**43.5%**

Table 5. Mean Ratio of Positives to Total Number of Videos for Pair-concept Queries

Dataset	IACC.2.B	IACC.2.C	IACC.2.B + C
Mean Ratio	**0.0159**	0.0134	0.0106

Table 6. Time Complexities Comparison

Method	Time Complexities
PicSOM	$O(1)$
Co-occurrence	$O(n)$
Fast0tag	$O(n)$
Rank-SVM	$O(n^2)$
Tagprob	$O(n^2)$
$f^{\mathrm{V}}_{\mathrm{corr\text{-}2}}$	$O(n^2)$
$f^{\mathrm{V}}_{\mathrm{corr\text{-}2}}$	$O(n^2)$
$f^{\mathrm{S}}_{\mathrm{latent}}$	$O(\|H\| \times n^2)$
$f^{\mathrm{VS}}_{\mathrm{latent}}$	$O(\|H\| \times \|Q\| \times n^2)$

Notice a drop in all the performances in Table 4. To explain this, we compute the ratio of an average number of positive of all queries and the total number of videos in each dataset (Table 5). Smaller ratio makes the retrieval harder, since there are less number of samples in training stage and also a higher drop even a single positive sample is lost in the testing stage.

4.6 Time Complexity

In Table 6, we show the time complexities comparison between different methods, where n is the number of training samples, $|H| \ll n$ is the number of shots in the video, and $|Q| \ll n$ is the number of concepts.

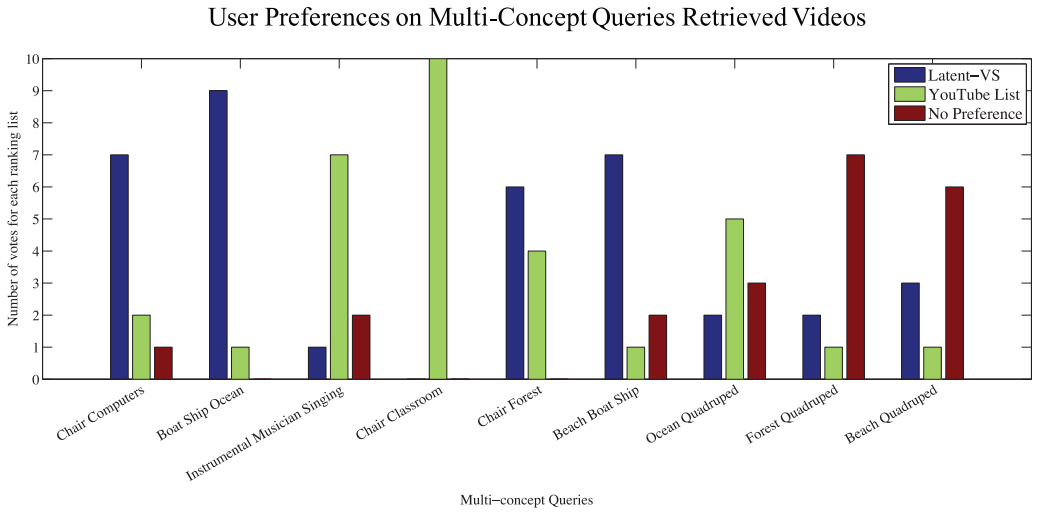

Fig. 4. The number of user votes for different ranking lists and queries. Users preferred the ranking lists by our algorithm to Youtube's default ranking lists for six out of nine queries.

4.7 User Study on YouTube Search Engine

We build a new dataset using 230 videos gathered from YouTube. These videos are retrieved using nine pair-concept queries shown in Table 1. Twenty to 30 videos for each of the queries are downloaded and their actual ranks in YouTube are saved. Shots are also built by clipping videos into 2s shots. We apply our method for each query and get a new ranking list. Ten randomly selected users are asked to compare these two ranking lists. One of them is YouTube original list and the other is from our method, and we ask users to choose the best list for corresponding query search, just based on visual cues. User voting is completely blind and participants do not have access to other participants' decisions. Even though our detectors are trained on the training set explained in Section 4.1.1, which is different from YouTube videos, in Figure 4, we show for some queries, users mostly like the re-ranked list using our method. YouTube ranking has a bias toward metadata coming with the videos and our method can retrieve videos containing the actual concepts instead.

4.8 Qualitative Analyses

We show some videos and their ranks using $f^{\mathrm{V}}_{\mathrm{avg}}$ and also after applying $f^{\mathrm{VS}}_{\mathrm{latent}}$ in Figure 5. In general, our approach works especially well under the following two scenarios. One is when the concepts appear in different video shots, and the other is when the concepts are not very relevant and more unlikely to happen in the same video (e.g., bridge, chair, and highway).

We also show how our method can correctly rank videos for a given query. In Figure 6, we show top five videos retrieved by our method for one of the pair-concept queries and a comparison with the average baseline. The top-ranked videos are more important especially for common users searches, since people tend to find a proper video at the top of their searching list result.

5 CONCLUSION

In this work, we have introduced a new baseline for multi-concept-based video retrieval. Our method uses a principled model that can be integrated and leveraged from other existing

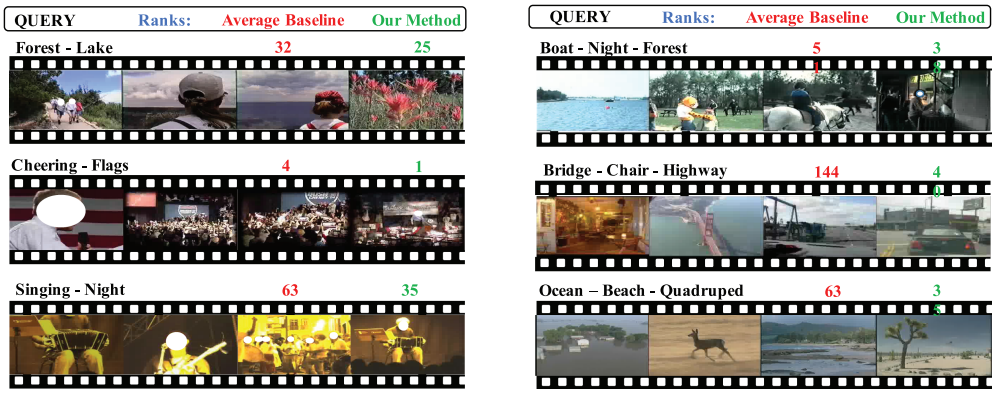

Fig. 5. Some examples of queries and the rank of one video containing all the concepts. We show two numbers for each video, the red one is the rank of the video using the Average baseline and the green one is the rank of the video using our method. We show queries with pair concepts and triplet concepts on the left and right columns, respectively. In all of the provided examples, the ranking value has been improved using our method.

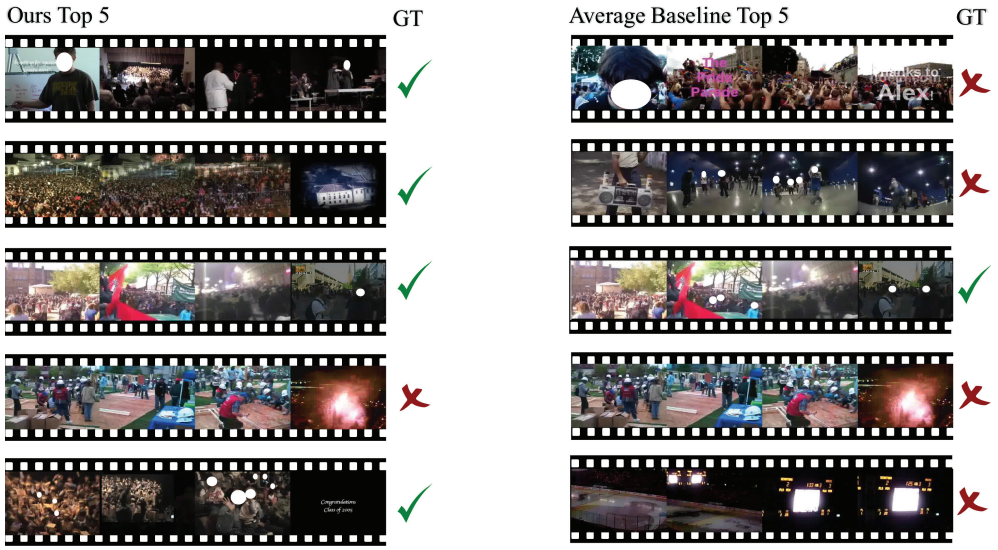

Fig. 6. Top-ranked videos for the query "Cheering-Night Time." The left panel shows the first top five videos ranked by our method and the right panel shows the same for the average baseline method. We also show which of the videos are tagged as positive in ground truth. Note that a positive video must have both concepts included.

methods of video/image retrieval. We have designed a multi-latent variable scoring function that can deal with noisy and incomplete annotations of large datasets, while it can learn both inter- and intra-shots dependencies of concepts. We show a technique named 0-1 loss-based early stopping criterion, which can make the training process more robust to outlier data. Our extensive experiments show our model superiority over other methods. As multi-concept-based video retrieval plays a central role in video retrieval systems, we expect our model to advance the state-of-the-art research in video retrieval.

REFERENCES

Tim Althoff, Hyun Oh Song, and Trevor Darrell. 2012. Detection bank: An object detection-based video representation for multimedia event recognition. In *Proceedings of the ACM Conference on Multimedia*.

Y. Alp Aslandogan and Clement T. Yu. 1999. Techniques and systems for image and video retrieval. *Knowl. Data Eng.* 11, 1 (1999), 56–63.

Shayan Assari, Amir Zamir, and Mubarak Shah. 2014. Video classification using semantic concept co-occurrences. In *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*.

Yusuf Aytar, Mubarak Shah, and Jiebo Luo. 2008. Utilizing semantic word similarity measures for video retrieval. In *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'08)*.

Murray Campbell, Alexander Haubold, Ming Liu, Apostol Natsev, John R. Smith, Jelena Tesic, Lexing Xie, Rong Yan, and Jun Yang. 2007. IBM research TRECVID-2007 video retrieval system. In *Proceedings of the NIST TRECVID Workshop*.

Shih-Fu Chang, Dan Ellis, Wei Jiang, Keansub Lee, Akira Yanagawa, Alexander C. Loui, and Jiebo Luo. 2007. Large-scale multimodal semantic concept detection for consumer video. In *Proceedings of the Workshop on Multimedia Information Retrieval*. ACM.

Shih-Fu Chang, Winston Hsu, Lyndon Kennedy, Lexing Xie, Akira Yanagawa, Eric Zavesky, and Dong-Qing Zhang. 2005. Columbia university trecvid-2005 video search and high-level feature extraction. In *Proceedings of the NIST TRECVID Workshop*.

Olivier Chapelle and S. Sathiya Keerthi. 2010. Efficient algorithms for ranking with SVMs. *Info. Retriev.* 13, 3 (2010), 201–215.

Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. 2014. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. *arXiv Preprint* (2014).

Afshin Dehghan, Mahdi M. Kalayeh, Yang Zhang, Haroon Idrees, Yicong Tian, Amir Mazaheri, Mubarak Shah, Jingen Liu, and Hui Cheng. 2014. UCF-CRCV at TRECVID 2014: Semantic indexing. In *Proceedings of the NIST TRECVID Workshop*.

Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. 2009. Describing objects by their attributes. In *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*.

David Grangier and Samy Bengio. 2008. A discriminative kernel-based approach to rank images from text queries. *Pattern Anal. Mach. Intell.* 30, 8 (2008), 1371–1384.

Matthieu Guillaumin, Thomas Mensink, Jakob Verbeek, and Cordelia Schmid. 2009. Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *Proceedings of the International Conference on Computer Vision (ICCV'09)*.

Alexander Haubold and Apostol Natsev. 2008. Web-based information content and its application to concept-based video retrieval. In *Proceedings of the International Conference on Content-based Image and Video Retrieval*. ACM.

Alexander Hauptmann, Rong Yan, and Wei-Hao Lin. 2007. How many high-level concepts will fill the semantic gap in news video retrieval? In *Proceedings of the ACM International Conference on Image and Video Retrieval*.

Ralf Herbrich, Thore Graepel, and Klaus Obermayer. 1999. Large margin rank boundaries for ordinal regression. *Adv. Neural Info. Process. Syst.* (1999), 115–132.

Rui Hou, Amir Roshan Zamir, Rahul Sukthankar, and Mubarak Shah. 2014. DaMN–Discriminative and mutually nearest: Exploiting pairwise category proximity for video action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV'14)*.

Weiming Hu, Nianhua Xie, Li Li, Xianglin Zeng, and Stephen Maybank. 2011. A survey on visual content-based video indexing and retrieval. *Syst., Man, and Cybernet., Part C: Appl. Rev.* 41, 6 (2011), 797–819.

Cisco Visual Networking Index. 2013. The zettabyte era–trends and analysis. *Cisco White Paper* (2013).

Cisco Visual Networking Index. 2014. Global mobile data traffic forecast. Retrieved from http://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white_paper_c11-520862.html.

Satoru Ishikawa, Markus Koskela, Mats Sjöberg, Jorma Laaksonen, Erkki Oja, Ehsan Amid, Kalle Palomäki, Annamaria Mesaros, and Mikko Kurimo. 2013. Picsom experiments in TRECVID 2013. In *Proceedings of the NIST TRECVID Workshop*.

Giridharan Iyengar and Harriet J. Nock. 2003. Discriminative model fusion for semantic concept detection and annotation in video. In *Proceedings of the ACM Conference on Multimedia*.

Kalervo Järvelin, Susan L. Price, Lois M. L. Delcambre, and Marianne Lykke Nielsen. 2008. Discounted cumulated gain-based evaluation of multiple-query IR sessions. In *Advances in Information Retrieval*. Springer, 4–15.

Lu Jiang, Shoou-I Yu, Deyu Meng, Teruko Mitamura, and Alexander G. Hauptmann. 2015. Bridging the ultimate semantic gap: A semantic search engine for internet videos. In *Proceedings of the ACM Conference on Multimedia*.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proceedings of the ACM SIGKDD Conference*.

Lyndon Kennedy, Shih-Fu Chang, and Apostol Natsev. 2008. Query-adaptive fusion for multimodal search. *Proc. IEEE* 96, 4 (2008), 567–588.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling. 2014. Attribute-based classification for zero-shot visual object categorization. *Pattern Anal. Mach. Intell.* 36, 3 (2014), 453–465.

Tian Lan, Weilong Yang, Yang Wang, and Greg Mori. 2012. Image retrieval with structured object queries using latent ranking svm. In *Proceedings of the European Conference on Computer Vision (ECCV'12)*.

Li-Jia Li, Hao Su, Li Fei-Fei, and Eric P. Xing. 2010. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in Neural Information Processing Systems*. 1378–1386.

Quannan Li, Jiajun Wu, and Zhuowen Tu. 2013. Harvesting mid-level visual concepts from large-scale internet images. In *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'13)*.

Xirong Li, Dong Wang, Jianmin Li, and Bo Zhang. 2007. Video search in concept subspace: A text-like paradigm. In *Proceedings of the ACM International Conference on Image and Video Retrieval*.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision* 60, 2 (2004), 91–110.

Amir Mazaheri, M. Kalayeh, Haroon Idrees, and Mubarak Shah. 2015. Ucf-crcv at trecvid2015: Semantic indexing. In *NIST TRECVID Workshop*.

Kieran McDonald and Alan F. Smeaton. 2005. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Image and Video Retrieval*. Springer, 61–70.

Apostol Paul Natsev, Alexander Haubold, Jelena Tešić, Lexing Xie, and Rong Yan. 2007. Semantic concept-based query expansion and re-ranking for multimedia retrieval. In *Proceedings of the ACM Conference on Multimedia*.

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

Paul Over, George Awad, Martial Michel, Jonathan Fiscus, Wessel Kraaij, Alan F. Smeaton, Georges Quéenot, and Roeland Ordelman. 2015. TRECVID 2015—An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of the NIST TRECVID Workshop*. NIST.

Paul Over, Georges Awad, Martial Michel, Johnatan Fiscus, Greg Sanders, Wessel Kraaij, Alan F. Smeaton, and Georges Quénot. 2014. TRECVID 2014- An overview of the goals. *Proceedings of the NIST TRECVID Workshop*.

O. Paul, G. Awad, M. Michel, J. Fiscus, W. Kraaij, A. F. Smeaton, and G. Quéenot. 2011. Trecvid 2011—An overview of the goals, tasks, data, evaluation mechanisms and metrics. *Proceedings of the NIST TRECVID Workshop*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

Florent Perronnin, Jorge Sánchez, and Thomas Mensink. 2010. Improving the Fisher kernel for large-scale image classification. *Proceedings of the European Conference on Computer Vision (ECCV'10)*.

James Petterson and Tibério S. Caetano. 2010. Reverse multi-label learning. In *Advances in Neural Information Processing Systems*. 1912–1920.

Wei Ping, Qiang Liu, and Alexander Ihler. 2014. Marginal structured SVM with hidden variables. *arXiv Preprint* (2014).

Guo-Jun Qi, Xian-Sheng Hua, Yong Rui, Jinhui Tang, Tao Mei, and Hong-Jiang Zhang. 2007. Correlative multi-label video annotation. In *Proceedings of the ACM Conference on Multimedia*.

Sreemanananth Sadanand and Jason J. Corso. 2012. Action bank: A high-level representation of activity in video. In *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*.

Amnon Shashua and Anat Levin. 2002. Ranking with large margin principle: Two approaches. In *Advances in Neural Information Processing Systems*. 937–944.

Behjat Siddiquie, Rogerio S. Feris, and Larry S. Davis. 2011. Image ranking and retrieval based on multi-attribute queries. In *Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'11)*.

Alan F. Smeaton. 2007. Techniques used and open challenges to the analysis, indexing and retrieval of digital video. *Info. Syst.* 32, 4 (2007), 545–559.

Alan F. Smeaton, Paul Over, and Wessel Kraaij. 2009. High-level feature detection from video in TRECVid: A 5-year retrospective of achievements. In *Multimedia Content Analysis*. Springer, 1–24.

C. G. M. Snoek, K. E. A. van de Sande, D. Fontijne, A. Habibian, M. Jain, S. Kordumova, Z. Li, M. Mazloom, S. L. Pintea, R. Tao, et al. 2013. MediaMill at TRECVID 2013: Searching concepts, objects, instances and events in video. In *Proceedings of the NIST TRECVID Workshop*.

Cees G. M. Snoek, Bouke Huurnink, Laura Hollink, Maarten De Rijke, Guus Schreiber, and Marcel Worring. 2007. Adding semantics to detectors for video retrieval. *Multimedia* 9, 5 (2007), 975–986.

Cees G. M. Snoek and Marcel Worring. 2008. Concept-based video retrieval. *Found. Trends Info. Retrieval* 2, 4 (2008), 215–322.

Cees G. M. Snoek, Marcel Worring, Jan C. Van Gemert, Jan-Mark Geusebroek, and Arnold W. M. Smeulders. 2006. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of the ACM Conference on Multimedia*.

Lorenzo Torresani, Martin Szummer, and Andrew Fitzgibbon. 2010. Efficient object category recognition using classemes. In *Proceedings of the European Conference on Computer Vision (ECCV'10)*.

Ioannis Tsochantaridis, Thomas Hofmann, Thorsten Joachims, and Yasemin Altun. 2004. Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the International Conference on Machine Learning*. ACM.

Vladimir Vapnik. 2013. *The Nature of Statistical Learning Theory*. Springer Science & Business Media.

Andrea Vedaldi and Brian Fulkerson. 2010. VLFeat: An open and portable library of computer vision algorithms. In *Proceedings of the ACM Conference on Multimedia*.

Dong Wang, Xirong Li, Jianmin Li, and Bo Zhang. 2007. The importance of query-concept-mapping for automatic video retrieval. In *Proceedings of the ACM Conference on Multimedia*.

Heng Wang and Cordelia Schmid. 2013. Action recognition with improved trajectories. In *Proceedings of the International Conference on Computer Vision (ICCV'13)*.

Xiao-Yong Wei, Chong-Wah Ngo, and Yu-Gang Jiang. 2008. Selection of concept detectors for video search by ontology-enriched semantic spaces. *Multimedia* 10, 6 (2008), 1085–1096.

Jun Wu and Marcel Worring. 2012. Efficient genre-specific semantic video indexing. *IEEE Trans. Multimedia* 14, 2 (2012), 291–302.

Rong Yan and Alexander G. Hauptmann. 2003. The combination limit in multimedia retrieval. In *Proceedings of the ACM Conference on Multimedia*.

Yang Yang and Mubarak Shah. 2012. Complex events detection using data-driven concepts. In *Proceedings of the European Conference on Computer Vision (ECCV'12)*.

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

Emine Yilmaz and Javed A. Aslam. 2006. Inferred ap: Estimating average precision with incomplete judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management (CIKM'06)*. 102–111.

Chun-Nam John Yu and Thorsten Joachims. 2009. Learning structural SVMs with latent variables. In *Proceedings of the International Conference on Machine Learning*. ACM.

Felix X. Yu, Rongrong Ji, Ming-Hen Tsai, Guangnan Ye, and Shih-Fu Chang. 2012. Weak attributes for large-scale image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*.

Yang Zhang, Boqing Gong, and Mubarak Shah. 2016. Fast zero-shot image tagging. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16)*. IEEE, 5985–5994.

Received May 2017; revised October 2017; accepted December 2017