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

Amir Mazaheri*, Boqing Gong†, Mubarak Shah ‡
Center for Research in Computer Vision(CRCV)
University of Central Florida
Orlando, USA
*Email: * amirmazaheri@knights.ucf.edu † bgong@crcv.ucf.edu ‡ shah@crcv.ucf.edu*

***Abstract*—Effective and efficient video retrieval has become a pressing need in the "big video" era and how to deal with multi-concept queries is a central component. The objective of this work is to provide a principled model for calculating the ranking scores of video in response to multiple concepts. However, it has been long overlooked and simply implemented by weighted averaging the corresponding concept detectors' scores. Our approach, which can be considered as a latent ranking SVM, integrates the advantages of various recent works on text and image retrieval, such as choosing ranking over structured prediction and modeling inter-dependencies between querying concepts and the others. Videos consist of shots and we use latent variables to account for the mutually complementary cues within and across shots. We introduce a simple and effective way to make our model robust to outliers and scarce data. Our approach gives rise to 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 seen queries used for training.**

***Keywords*-video retrieval; multi-concept retrieval; structural learning**

I. Introduction

Video data is explosively growing from surveillance, health care, and personal mobile phones to name a few sources. 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 [1], the most efforts have been mainly spent on *(multi-)concept* based video retrieval [2], 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. We refer the readers to the survey papers [2], [1] and the annual TRECVID workshops [3] 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, from 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

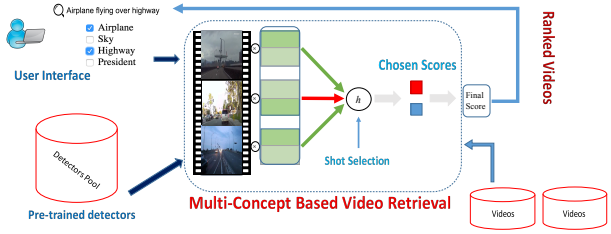

Figure 1. How to calculate the ranking scores of video in response to one or more concepts, is the central component in many video retrieval systems. It takes as input the multi-concept queries and then returns a ranked list of videos. The multiple concepts in a query could be directly supplied by the users or inferred by the systems from the users' text queries.

on the Web [4]. Correspondingly, a plethora of works has been devoted to concept detectors [5], [6], [7], [8].

See Figure 1. Users can directly select concepts from a checklist to compose the queries. In this paper, 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 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 [9], where the weights are either uniform or in some systems derived from the similarities between an open-vocabulary user query and the selected concepts.

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 systems based on user queries.

Our approach, which can be considered as a latent ranking SVM [10], integrates different advantages of the recent works on text and multi-attribute based image retrieval. Particularly, we model the video retrieval as a ranking problem following [11], as opposed to the structured prediction problem used in [12], [13], in order to harvest better efficiency and larger modeling capacity to accommodate a latent variable. The latent variable is used for us 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 [12], [13] that the inter-dependencies between both selected and unselected concepts 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.

II. Related Work

Concept-based video retrieval has been widely recognized as the promising direction for filling in the semantic gap in video retrieval [2], [14], since concepts summarize the low-level visual cues to intermediate representations which humans can read and understand. After mapping user queries to one or more concepts [15], [16], the "retrieval" component of a system searches the database and returns the videos relevant to those concepts [17]. This is often conducted by manually defined similarities and heuristic fusion techniques [18], [9] to relate the querying concepts with the concept detections from the videos. In contrast, this work provides a principled model to automatically learn and rank the videos given multiple concepts. The concepts can take the forms of action bank [19], image cell based detections [20], classemes [21], sentiment concepts [22], events [8], and so on. Whereas many concept detectors are trained from manually labeled datasets [23], some other work harvests detectors from the noisy Web data [24]. Almost all the existing concept detectors can be conveniently integrated with our multi-concept based video retrieval model.

Image retrieval and ranking using multiple attributes [12], [13] are the most relevant problem to our multi-concept based video retrieval. However, due to the vast number of video shots in a database, the structural SVM [25], [26] model used in [12], [13] becomes intractable in our experiments. Instead, we develop a simpler ranking model for our retrieval problem.

There is a pile of works on learning to rank using the large-margin principle [27], [28], [11], [10]. The conventional ranking SVMs [29], [27] only learn the ranking function for one query, while we learn the model parameters not only for more than one training queries but also to generalize them to previously unseen queries.

III. Approach

Denote by $\mathcal{Q}$ all the concepts offered by the system for the users to compose queries, by $\mathcal{V}$ all the videos in the the database, and by $R(Q) \subset \mathcal{V}$ the videos that are related to a multi-concept query $Q \subset \mathcal{Q}$. Accordingly, the retrieval model should possess some mechanism to differentiate the groundtruth subset $R(Q)$ from any other subsets of the videos in the database and tell that

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

given the querying concepts in Q. Directly modeling this notion gives rise to a structured prediction model presented in [12] and strengthened in [13]. Unfortunately, it suffers from high computation costs due to the exponential number $2^{|\mathcal{V}|}$ of distinct subsets of $\mathcal{V}$. Instead, we relax the retrieval to a ranking problem as in [11] and managed to accommodate multiple latent variables to tackle the shot-level detections. In particular, the rigorous criterion (eq. (1)) for retrieval is replaced by a less constrained ranking criterion,

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

where V_i and V_j are a pair of videos in the database $\mathcal{V}$.

Comparing eq. (1) with eq. (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 $V \in \mathcal{V}$. We use the following ranking model in this work $\mathcal{F} : \mathcal{Q} \times \mathcal{V} \mapsto \mathbb{R}$,

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

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

Given a multi-concept query Q, we simply rank the videos by $\mathcal{F}$ and return the top portion of the ranking list to the user. In order to train the model, we employ the strategy used in ranking SVM [29], [27] and arrive at the following:

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

where $\mathcal{N}(Q)$ is the collection of all the pairs of videos V_i and V_j in eq. (2) for the query Q, and $L(x) \geq 0$ is a loss function. The loss function will impose some amount of penalty when the ranking scores of a pair of videos violate the ranking constraint of eq. (2).

We exploit two types of loss functions in this work, the hinge loss $L_{\text{hinge}}(x) = \max(x + 1, 0)$ 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;* we instead use it to define a novel early stopping criterion when we solve the problem with hinge loss by sub-gradient descent. Namely, the program stops when the change of the objective function value, 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. 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 [11], which motivates our ranking model.

A. Video-level concept detection

To quickly respond to the users' queries, it is often the case that the concept detection results $\boldsymbol{\phi}(V)$ over each video $V \in \mathcal{V}$ are computed off-line and then stored somewhere. 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 this vector, the video-level concept detection results, to design the scoring functions $f(q), q \in Q \subset \mathcal{Q}$. We start from the weighted average which prevails in the existing video retrieval works.

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 (eq. (3)). A common practice to rank the videos given a multi-concept query Q, is by the weighted average of the corresponding concept detection confidences:

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

where the weights $\boldsymbol{\Theta}_{qq}, q \in Q$ could be the similarities between the concepts in Q and an open-vocabulary user query [15], [16]. We only study the uniform weights $\boldsymbol{\Theta}_{qq} = 1$ in this work without loss of generality.

Note that this weighted average fails to model either 1) the correlations between the concepts in Q or 2) the correlations between Q and the remaining unselected concepts in $\mathcal{Q}$. To see this point more clearly, we denote by $\mathbf{1}^q \in \mathbb{R}^{|\mathcal{Q}|}$ the one-hot vector taking the value 1 at the q-th entry and zeros else. The rightmost of eq. (5) thus follows. Further, the model parameters $\boldsymbol{\Theta} = (\mathbf{1}^1, \mathbf{1}^2, \cdots, \mathbf{1}^{|\mathcal{Q}|})^T = I \in \mathbb{R}^{|\mathcal{Q}| \times |\mathcal{Q}|}$ actually correspond to 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 (eq. (5)).

B. Shot-level concept detection

In practice many concept detectors actually take the video shots, or even frames, as the input [8], [19], [20]. Suppose for a video V in the database we have partitioned it into H shots. A video retrieval system can then pre-compute and store the concept detection results $\boldsymbol{\phi}^h \in \mathbb{R}^{|\mathcal{Q}|}, h = 1, \cdots, \mathsf{H}$ for all the concepts $\mathcal{Q}$ over the shots of the video. Compared to the video-level concept detections, the shot-level detections provide more insights and finer-grained information about the video database. We thus propose some new form of scoring function to take advantage of such detection results. 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.

$$f^{\text{VS}}_{\text{latent}}(q) = \max_{h \in \{1, \cdots, \mathsf{H}\}} \left\langle \boldsymbol{\theta}^q, \boldsymbol{\phi}^h \right\rangle + \sum_{p \in \mathcal{Q}} \max_{g \in \{1 \cdots, \mathsf{H}\}} \upsilon^q_p \phi^g_p, \tag{6}$$

where the model parameters $\boldsymbol{\theta}^q \in \mathbb{R}^{|\mathcal{Q}|}$, which correspond to the concept $q \in Q \subset \mathcal{Q}$, count the contributions to q from all the concepts within the shot, which will be selected by the latent variable $h \in \{1, 2, \cdots, \mathsf{H}\}$. Also, $\max_g \upsilon^q_p \phi^g_p$ 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 $f^{\text{VS}}_{\text{latent}}(q)$. The first term is robust to the concepts which are negatively correlated with q. 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. The second terms strengthen the detection score of concept q from some positively correlated concepts in the video since it is more likely for two positive concepts to happen in two different shots rather than same shots. The model parameters $\boldsymbol{\theta}^q$ and $\boldsymbol{\upsilon}^q$ are learned by solving eq. (4) with sub-gradient descent. Details are shown as follow.

Optimization: We solve for the model parameters by (sub-)gradient descent. As discussed in Section 3.1, the loss function L is non-zero on the pairs $\{(V_i, V_j)\}$, for each of which the negative video V_j has higher ranking score than the positive V_i. As a result, we get the gradients on those pairs and for the other pairs the gradients are simply zero.

Denoting by

$$\mathcal{S}_j = \frac{\partial L}{\partial \mathcal{F}(Q, V_j)} \times \frac{\partial \mathcal{F}(Q, V_j)}{\partial \boldsymbol{\Theta}}, \tag{7}$$

we thus have the overall gradients of eq. (4) by

$$\sum_Q \frac{1}{|\mathcal{N}(Q)|} \sum_{(i,j) \in \mathcal{N}(Q)} (\mathcal{S}_j - \mathcal{S}_i)\,. \tag{8}$$

Note that the model parameters $\boldsymbol{\Theta}$ consist of two parts $(\boldsymbol{\theta}, \boldsymbol{\upsilon})$, corresponding to the two terms of $f^{\text{VS}}_{\text{latent}}$ (cf. eq. (6)), respectively. We compute the gradients with respect to the first part $\boldsymbol{\theta}$ using the $softmax$ derivation to approximate a smooth gradients, as suggested by [30]:

$$\frac{\partial \mathcal{F}(Q, V_j)}{\partial \boldsymbol{\theta}} = \sum_{h \in \{1, 2, \cdots, \mathsf{H}\}} \frac{e^{\langle \boldsymbol{\theta}^q, \boldsymbol{\phi}^h \rangle} \boldsymbol{\phi}^h}{\sum_{j \in \{1, 2, \cdots, \mathsf{H}\}} e^{\langle \boldsymbol{\theta}^q, \boldsymbol{\phi}^j \rangle}}. \tag{9}$$

We write out the gradients with respect to the second part $\boldsymbol{\upsilon}$ over different dimensions of $\boldsymbol{\upsilon}$. Recall that he second term of $f^{\text{VS}}_{\text{latent}}$ (cf. eq. (6)), $\max_g \upsilon^q_p \phi^g_p$, max-pools over all the shots of a video for each single concept. As a result, we have the following:

$$\frac{\partial \mathcal{F}}{\upsilon^q_p} = \phi^{g*}_p, \quad p = 1, 2, \cdots, \mathsf{H},\ q = 1, \cdots, \mathsf{H} \tag{10}$$

where $g*$ is determined by $g* \leftarrow \max_g v_p^q \phi_p^g$.

IV. Experimental Results

Our experiments depend on two separate datasets respectively for video retrieval and training the concept detectors. We further exploit three sets of multi-concept queries. Two sets consists of 50 queries in the form of concept pairs, one for training and testing and the other just for testing. The other sets contain 50 triplets concepts queries to be used just in testing. We train our model using only the first set of queries on the training set, and then test it by all three sets of queries on the test set.

A. Datasets

Our approach is mainly tested over the IACC.2.B dataset which is the test set used in the Semantic Indexing (SIN) task of TRECVID 2014 [3]. We randomly split IACC.2.B into a training set of 712 videos, a validation set of 474 videos, and a test set of 1,185 videos. From the $\binom{30}{2}$ possible pairs of concepts, we select 50 pairs as the first set of queries in our experiments. For each query, we consider that a video is related when each concept in the query has at least one positive shot in the video. 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. Additionally, we also build the second set of queries with 50 concept triples. There are on average 24 related videos to a concept-triplet query. Note that the more concepts a query comprises, the more challenging the retrieval task is due to the smaller number of related videos. We also test our main approach on IACC.2.C dataset which is the test set for SIN task of TRECVID 2015 [14], with similar settings as a correctness proof of our approach.

B. Concept detectors

It has been an active research area to learn robust concept detectors for videos [5], [8], [6]. Virtually all kinds of concept detectors can be employed in our retrieval model, as long as they output the shot or video level detection confidences. We train our own detectors following the practice of [7]. In particular, we train 60 independent detectors from the training data (IACC.1.tv10.training, IACC.1.A&B&C) of the TRECVID 2014 SIN task [3].

C. Practical considerations in implementation

In our implementation, we add $\sum_{q\in\mathcal{Q}} \lambda \left\|\boldsymbol{\theta}^q\right\|_2^2 + \gamma \left\|\boldsymbol{v}^q\right\|_2^2$ to regularize the optimization problem in eq. (4) and tune the hyper-parameters λ and γ using the validation set. Note that $\gamma = 0$ except for the scoring function $f^{\text{VS}}_{\text{latent}}(q)$. For the shot-level concept detections, we impose symmetric constraints over the model parameters (i.e., $\boldsymbol{\theta}_p^q = \boldsymbol{\theta}_q^p$). When we train the model with latent variable h (cf. $f^{\text{S}}_{\text{latent}}$ and $f^{\text{VS}}_{\text{latent}}$), we remove the shots without annotations from negative training videos. Both $\boldsymbol{\theta}^q$ and $\boldsymbol{v}^q$ are initialized by one-shot vectors. The time complexity of training a pairwise ranking loss function is $O(n^2)$ [31].

D. Comparison results

In Table I, we report the comparison results of different scoring functions that account for different types of concept detectors, evaluated using Normalized Discounted Cumulative Gain NDCG [35]. Given a ranking list for query Q, NDCG@k is calculated by:

$$\text{NDCG@}k = \frac{1}{Z}\sum_{j=1}^{k}\frac{G[j]}{1+\log_2 j}, \tag{11}$$

where j is the rank of a video and $G[j] = \text{rel}(j)^2$ with $\text{rel}(j)$ being the number of positive concepts shared by that video and the query Q. The partition Z is calculated from the ideal ranking list such that any NDCG@k value is normalized between 0 and 1. We shall report the results at $k = 5, 10, \cdots, 50$ in the following experiments.

Following the common practice in the existing concept-based video retrieval systems, we empirically test a variety of fusion methods [17], [18], [34] as the (old) baselines—our approach offers a new simple yet more advanced baseline for the 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 [17]. We thus only show the results of $f^{\text{V}}_{\text{avg}}$ and the second best, PicSOM [34], in the rows tagged by "Common practice" and "PicSOM 2013", respectively, in Table I. Also, we used a another common technique as explained in [33] 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 [28] and TagProp [32] in the table. Both takes as input the video-level representations. TagProp is a state-of-the-art image tagging algorithm. We reported the best results for each after parameter tuning.

There are $|\mathcal{Q}| = 30$ concepts labeled for our video database $\mathcal{V}$, which is drawn from the IACC.2.B dataset. All queries are constructed from these concepts such that we have the groundtruth ranking list for evaluation. We first see the video retrieval results in using 30 concepts of Table I. Our model improves the $f^{\text{V}}_{\text{avg}}$ by a significant margin.

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 III). The bottom half of Table I shows the results corresponding to 60 concept detectors. We see that 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 observation in [13]. Indeed, the inter-dependences of more concepts may provide better cues

Table I
COMPARISON RESULTS OF DIFFERENT METHODS IN PAIR-CONCEPT BASED VIDEO RETRIEVAL.

Functions		NDCG@5	@10	@15	@20	@25	@30	@35	@40	@45	@50	Mean
Baselines												
Common practice	$f^{\mathrm{V}}_{\mathrm{avg}}$	0.626	0.571	0.556	0.561	0.575	0.588	0.597	0.610	0.620	0.626	0.593
TagProp [32]		0.300	0.273	0.256	0.268	0.277	0.286	0.294	0.301	0.308	0.314	0.288
Rank-SVM [29]		0.579	0.529	0.522	0.526	0.543	0.554	0.565	0.568	0.577	0.581	0.555
Co-occurrence [33]		0.594	0.507	0.486	0.495	0.518	0.534	0.549	0.556	0.564	0.574	0.538
PicSOM 2013 [34]		0.630	0.571	0.555	0.559	0.573	0.581	0.592	0.605	0.615	0.621	0.590
$\|\mathcal{Q}\|$= **30 concepts**												
Shot-level (0-1 loss)	$f^{\mathrm{VS}}_{\mathrm{latent}}$	0.629	0.588	0.583	0.600	0.618	0.631	0.647	0.654	0.662	0.671	0.628
$\|\mathcal{Q}\|$= **60 concepts**												
Shot-level (Hinge loss)	$f^{\mathrm{VS}}_{\mathrm{latent}}$	0.654	0.592	0.577	0.591	0.609	0.629	0.643	0.653	0.667	0.674	0.633
Shot-level (0-1 loss)	$f^{\mathrm{VS}}_{\mathrm{latent}}$	**0.698**	**0.638**	**0.617**	**0.609**	**0.630**	**0.641**	**0.654**	**0.664**	**0.668**	**0.674**	**0.649**

Table II
BASELINE AVERAGES OF NDCG@5-50 ON THREE DIFFERENT SETS AND THE MEAN RATIO OF POSITIVES TO TOTAL NUMBER OF VIDEOS

Dataset	IACC.2.B	IACC.2.C	IACC.2.B + C
TagProp	0.2888	0.241	0.123
Rank-SVM	0.555	0.264	0.395
Co-occurrence	0.538	0.444	0.302
$f^{\mathrm{V}}_{\mathrm{avg}}$	0.593	0.537	0.404
$f^{\mathrm{VS}}_{\mathrm{latent}}$	**0.649**	**0.569**	**0.435**
Mean Ratio	**0.0159**	0.0134	0.0106

for our scoring functions and make them more robust to the unreliable concept detection confidences.

E. Generalizing out of the training queries

After training our models using the pair-concept queries, we expect it to also generalize well to other multi-concept queries. We choose two sets of new queries for this experiment: (a) 50 new pair-concept queries, and (b) 50 new triple-concept queries. None of them are used to train our model. Figure 2 shows the retrieval results using our model. We can see our model with the shot-level scoring functions $f^{\mathrm{VS}}_{\mathrm{latent}}$ performs quite well upon the new queries. The results are not only significantly better than $f^{\mathrm{V}}_{\mathrm{avg}}$ but also comparable to those on the seen pair-concept queries (cf. Table I).

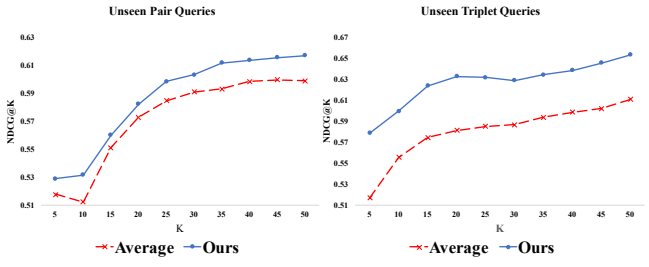

Figure 2. The video retrieval results using *previously unseen* queries: (a) pair-concept queries, and (b) triple-concept queries.

F. Extensive Experiments

We extended our experiments using IACC.2.C dataset by dividing it to training, validation and testing sets with similar settings as explained in Section 4.A. Also, by integrating it with IACC.2.B, we built one super dataset. Different experiments shown in Table II are various methods and all of them had a drop, however, in all of three cases we see an improvement after applying our method. By decreasing the ratio of positives to total number of samples in datasets, we see a drop in baseline performance. 50 queries with highest frequency of positives has been considered for each set.

V. CONCLUSION

We have developed a principled model for multi-concept based video retrieval. It integrates the advantages of several existing methods on text based and multi-attribute based image retrieval. In addition, we introduce latent variables and the 0-1 loss based early stopping criterion to model the temporal structures and the noisy labels in videos, respectively. Our experiments clearly verify the effectiveness of the proposed model for not only queries seen at training, but also previously unseen queries.

ACKNOWLEDGMENT

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

REFERENCES

[1] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, "A survey on visual content-based video indexing and retrieval," *Systems, Man, and Cybernetics, Part C: Applications and Reviews*, 2011.

[2] C. G. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, 2008.

[3] P. Over, J. Fiscus, G. Sanders, D. Joy, M. Michel, G. Awad, A. Smeaton, W. Kraaij, and G. Quénot, "Trecvid 2014- an overview of the goals," *NIST TRECVID Workshop*, 2014.

[4] A. Hauptmann, R. Yan, and W.-H. Lin, "How many high-level concepts will fill the semantic gap in news video retrieval?" in *ICIVR*. ACM, 2007.

[5] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang, "Correlative multi-label video annotation," in *ACM Multimedia*, 2007.

[6] C. Snoek, K. Sande, D. Fontijne, S. Cappallo, J. Gemert, A. Habibian, T. Mensink, P. Mettes, R. Tao, D. Koelma *et al.*, "Mediamill at trecvid 2013: Searching concepts, objects, instances and events in video," in *NIST TRECVID Workshop*, 2013.

[7] A. Mazaheri, M. Kalayeh, H. Idrees, and M. Shah, "Ucf-crcv at trecvid 2015: Semantic indexing," in *NIST TRECVID Workshop*, 2015.

[8] G. Ye, Y. Li, H. Xu, D. Liu, and S.-F. Chang, "Eventnet: A large scale structured concept library for complex event detection in video," in *ACM Multimedia*, 2015.

[9] Y. Aytar, M. Shah, and J. Luo, "Utilizing semantic word similarity measures for video retrieval," in *CVPR*, 2008.

[10] T. Lan, W. Yang, Y. Wang, and G. Mori, "Image retrieval with structured object queries using latent ranking svm," in *ECCV*, 2012.

[11] D. Grangier and S. Bengio, "A discriminative kernel-based approach to rank images from text queries," *PAMI*, 2008.

[12] B. Siddiquie, R. S. Feris, and L. S. Davis, "Image ranking and retrieval based on multi-attribute queries," in *CVPR*, 2011.

[13] F. X. Yu, R. Ji, M.-H. Tsai, G. Ye, and S.-F. Chang, "Weak attributes for large-scale image retrieval," in *CVPR*, 2012.

[14] P. Over, J. Fiscus, G. Sanders, D. Joy, M. Michel, G. Awad, A. Smeaton, W. Kraaij, and G. Quénot, "Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics," in *NIST TRECVID Workshop*, 2015.

[15] D. Wang, X. Li, J. Li, and B. Zhang, "The importance of query-concept-mapping for automatic video retrieval," in *ACM Multimedia*, 2007.

[16] L. Kennedy, S.-F. Chang, and A. Natsev, "Query-adaptive fusion for multimodal search," *Proceedings of the IEEE*, 2008.

[17] M. Campbell, A. Haubold, M. Liu, A. Natsev, J. R. Smith, J. Tesic, L. Xie, R. Yan, and J. Yang, "Ibm research trecvid-2007 video retrieval system." in *NIST TRECVID Workshop*, 2007.

[18] R. Yan and A. G. Hauptmann, "The combination limit in multimedia retrieval," in *ACM Multimedia*, 2003.

[19] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *CVPR*, 2012.

[20] T. Althoff, H. O. Song, and T. Darrell, "Detection bank: an object detection based video representation for multimedia event recognition," in *ACM Multimedia*, 2012.

[21] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *ECCV*, 2010.

[22] T. Chen, D. Borth, T. Darrell, and S.-F. Chang, "Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks," *arXiv preprint*, 2014.

[23] C. H. Lampert, H. Nickisch, and S. Harmeling, "Attribute-based classification for zero-shot visual object categorization," *PAMI*, 2014.

[24] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *CVPR*, 2013.

[25] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, "Support vector machine learning for interdependent and structured output spaces," in *ICML*. ACM, 2004.

[26] C.-N. J. Yu and T. Joachims, "Learning structural svms with latent variables," in *ICML*. ACM, 2009.

[27] R. Herbrich, T. Graepel, and K. Obermayer, "Large margin rank boundaries for ordinal regression," *NIPS*, 1999.

[28] T. Joachims, "Optimizing search engines using clickthrough data," in *ACM SIGKDD*, 2002.

[29] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with svms," *Information Retrieval*, 2010.

[30] W. Ping, Q. Liu, and A. Ihler, "Marginal structured svm with hidden variables," *arXiv preprint*, 2014.

[31] F. L. Wauthier, M. I. Jordan, and N. Jojic, "Efficient ranking from pairwise comparisons." *ICML (3)*, vol. 28, pp. 109–117, 2013.

[32] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid, "Tagprop: Discriminative metric learning in nearest neighbor models for image auto-annotation," in *ICCV*, 2009.

[33] S. Assari, A. Zamir, and M. Shah, "Video classification using semantic concept co-occurrences," in *CVPR*, 2014.

[34] S. Ishikawa, M. Koskela, M. Sjöberg, J. Laaksonen, E. Oja, E. Amid, K. Palomäki, A. Mesaros, and M. Kurimo, "Picsom experiments in trecvid 2013," in *NIST TRECVID Workshop*, 2013.

[35] K. Järvelin, S. L. Price, L. M. Delcambre, and M. L. Nielsen, "Discounted cumulated gain based evaluation of multiple-query ir sessions," in *Advances in Information Retrieval*. Springer, 2008.