Online Reranking via Ordinal Informative Concepts for Context Fusion in Concept Detection and Video Search

Yi-Hsuan Yang, Winston H. Hsu, and Homer H. Chen, *Fellow, IEEE*

***Abstract*—To exploit the co-occurrence patterns of semantic concepts while keeping the simplicity of context fusion, a novel reranking approach is proposed in this paper. The approach, called ordinal reranking, adjusts the ranking of an initial search (or detection) list based on the co-occurrence patterns obtained by using ranking functions such as ListNet. Ranking functions are by nature more effective than classification-based reranking methods in mining ordinal relationships. In addition, the ordinal reranking is free of the *ad hoc* thresholding for noisy binary labels and requires no extra offline learning or training data. To select informative concepts for reranking, we also propose a new concept selection measurement, *wc-tf-idf*, which considers the underlying ordinal information of ranking lists and is thus more effective than the feature selection algorithms for classification. Being largely unsupervised, the reranking approach to context fusion can be applied equally well to concept detection and video search. While being extremely efficient, ordinal reranking outperforms existing methods by up to 40% in mean average precision (MAP) for the baseline text-based search and 12% for the baseline concept detection over TRECVID 2005 video search and concept detection benchmark.**

***Index Terms*—Context fusion, learning-to-rank, rerank, video concept detection, visual search, wc-tf-idf.**

I. Introduction

TO FACILITATE random access and semantic understanding of large-scale multimedia databases, image/video retrieval and semantic concept detection [3] have been an active research area, thanks to the continuing growth of home videos, photo collections, broadcast news videos, and media sharing in the emerging social networks.

Bridging the semantic gap—the chasm between raw data (signals) and high-level semantics (meanings)—is essential for exploiting the growing multimedia contents. Toward this goal, recent research has focused on building detectors for detecting *concepts* such as locations, objects, and people [3] using a predefined lexicon and a sufficient number of annotated examples. Once trained, these detectors can be used to semantically tag and index multimedia contents in a fully automatic fashion.

On the other hand, there has been a substantial body of work on *visual search*, whose goal is to find images or videos in response to queries that are unknown to the system. Current visual search solutions are mostly restricted to text-based approaches which process keyword queries against text tokens associated with the media, such as speech transcripts, captions, file names, etc. However, such textual information may not necessarily come with the image or video. It has been shown that the use of other modalities such as image and audio content improves text-based visual search [8], [17]–[19], but it requires multiple example images, which could be difficult for users to prepare. Additionally, it is observed that most users expect to search simply through a few keywords [9].

Being two extreme scenarios (supervised versus unsupervised), concept detection and visual search actually share a unified goal: finding videos or images meeting certain semantic information needs (or *target semantics*). Therefore, success in one scenario should benefit the other. Since concept detection is likely to provide high-level contexts,[1] one promising direction is to utilize auxiliary concepts to aide (supervised) concept detection and even (unsupervised) visual search. For example, since the detection accuracy of the concept *person* is high, we can use its detection result to help the detection of related, yet more difficult, concepts such as *people marching* or *crowd*. Likewise, since "Hu Jintao" is contextually related to concepts such as *government leader*, *office*, and *Asian people*, "Searching videos of Hu Jintao," could be easier by incorporating the detection results of these concepts.

The use of peripherally related concepts to refine detection or search of semantic targets is generally called *context fusion*.

Manuscript received September 15, 2008; revised March 21, 2009. First version published July 7, 2009; current version published December 1, 2009. This work was supported in part by National Taiwan University, under Contract NTU97R0062-04, and grants from the National Science Council of Taiwan, under Contracts NSC 97-2221-E-002-111-MY3 and NSC 96-2628-E-002-005-MY2. A preliminary version of this paper appears in Proc. ACM Multimedia 2008 [2]. This paper was recommended by Associate Editor R. C. Lancini.

Y.-H. Yang is with the Graduate Institute of Communication Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: affige@gmail.com).

W. H. Hsu is with the Graduate Institute of Networking and Multimedia and the Department of Computer Science and Information Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: winston@csie.ntu.edu.tw).

H. H. Chen is with the Graduate Institute of Communication Engineering, the Graduate Institute of Networking and Multimedia and the Department of Electrical Engineering, National Taiwan University, Taipei 10617, Taiwan (e-mail: homer@cc.ee.ntu.edu.tw).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2009.2026978

[1]The meanings of "context" are usually application-dependent [34]. Here, we refer to context as those attributes describing *who*, *where*, *when*, *what*, etc., shared by documents forming the recurrent patterns.

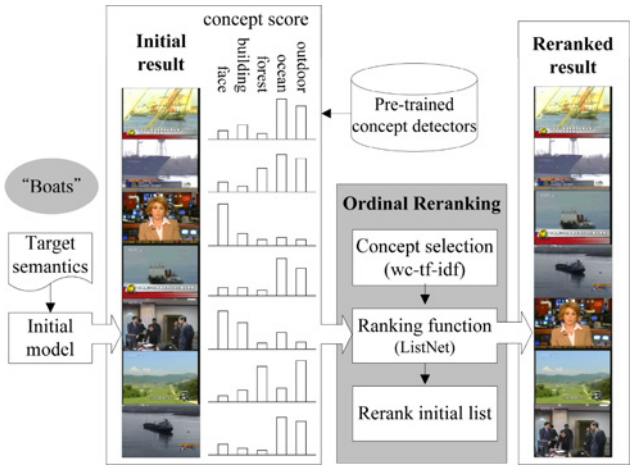

Fig. 1. Architecture of the proposed ordinal reranking framework for context fusion, with applications to video search and concept detection. The initial result from a text-only search model or a pre-trained concept detector is taken as an approximation of the target semantics. A ranking algorithm is then employed to mine the co-occurrence patterns of extracted features (i.e., concept detection scores [4]) to rerank the initial result.

It was first explored in prior work for concept detection [15]–[18] and then extended to video search under a *reranking* framework [5]–[7], which aims to leverage the co-occurrence patterns between target semantics and extracted features (high-level concept detectors or low-level visual features) to refine (by reordering) the result of an initial text-based search. A typical approach, referred to as *classification-based reranking* [7] in this paper, takes the higher rank and lower rank results of a baseline system as pseudopositive and pseudonegative examples to train a discriminative model and regards the normalized classification score for each object in the initial list as its reranked score. Since reranking is largely unsupervised, it can be applied equally well to context fusion in both concept detection and video search tasks. Salient performance gain over baseline methods has been reported in [7]. In addition, the simplicity of the keyword-based search paradigm is maintained under the reranking framework.

Though the classification-based reranking method has the advantage that existing classification methodologies can be directly applied, it is not free of problems. First, the formulation of learning as a minimization of classification errors neglects the underlying ordinal information of the initial list. Second, the classification-based reranking resorts to an *ad hoc* mechanism for determining the threshold for noisy binary labels. In addition, determining the pools of the pseudopositive and pseudonegative sets, which is vital to the system performance, is not straightforward.

In this paper, to exploit contextual information for concept detection and visual search, we propose a novel reranking method, called *ordinal reranking*, that employs ranking algorithms such as RankSVM [11] and ListNet [20] to learn the co-occurrence patterns between target semantics and features extracted from the initial list. Since the objective function is to minimize errors in object ranking, ordinal reranking is more effective and efficient for mining ordering information and free of the *ad hoc* thresholding problem.

Fig. 1 gives an illustrative example. A baseline model, which can be a text-based search model or a pre-trained concept detector, retrieves video shots (basic video retrieval units) that match the target semantics "Find shots of boats." Besides false positives (e.g., images with an anchorperson or crowds), there are still certain relevant shots ranked low due to the semantic gap or the lack of associated keyword annotations. From the noisy initial ranked list, ordinal reranking mines the co-occurrence patterns and identifies "ocean" and "outdoor" as the relevant concepts. Reranking is then made by reordering the shots with high search scores linearly fused by these relevant concepts.

We further investigate concept selection methods that automatically select informative concepts for reranking in an unsupervised fashion. Considering visual objects (video shots or images) as documents and concepts as visual terms, we improve the c-tf-idf (concept tf-idf[2]) measurement [10] by incorporating the ordinal information provided by the initial list. The new measurement, weighted c-tf-idf (wc-tf-idf), has promising performance and further improves reranking.

Besides being extremely efficient, ordinal reranking outperforms existing reranking methods and improves up to 40% in mean average precision (MAP)[3] for the baseline text-based search and 12% for the baseline concept detection, when evaluated on the TRECVID 2005 video search and concept detection benchmark [1].

In summary, the primary contributions of the paper include the following.

1) To our best knowledge, the proposed ordinal reranking method represents one of the first attempts that utilize ranking algorithms for reranking (Section IV). Because the underlying ordinal information is better exploited, ordinal reranking outperforms existing reranking methods in both effectiveness and efficiency.
2) As far as we know no feature selection measure has been designed specifically for reranking. We adapt the famous tf-idf measurement to the reranking framework by taking the ordering of the objects into account (Section V).
3) An extensive performance study including comparisons to existing reranking methods, parameter sensitivity test, and analysis of the result of concept selection, is conducted (Section VI).

II. Related Work

As shown in Table I, context fusion approaches can be categorized into "offline" and "online" methods. Offline methods use annotations of training data to discover the contextual information, while online methods approximate the initial result of a baseline system as pseudoground truth to further rerank the initial result. Below we briefly review some existing methods.

[2]"tf-idf" stands for term-frequency inverse-document-frequency [14].

[3]MAP: mean average precision, a performance metric used in TRECVID for concept detection and search [1].

TABLE I
COMPARISON OF CONTEXT FUSION APPROACHES

	Offline (Supervised)	Online (Unsupervised)	
		Classification-based reranking	Ordinal reranking
Learning strategy	Using annotations of training data to discover contextual information to rerank test data	Using initial result of a baseline system as pseudoground truth to learn to rerank the initial result	
		Higher rank (positive), lower rank (negative)	Maintaining ranking order
Advantage	–	Unsupervised, comparable performance to offline methods	More effective and efficient in mining ordinal information
Drawback	Not applicable to image/video search	*Ad hoc* thresholding, losing ordinal information	–
References	[7], [15]–[17], [21]–[27]	[6]–[8], [19], [28]	This paper

A. Offline Context Fusion

The context-based concept fusion approach has been explored in prior work for concept detection [15]–[17], [21]–[23]. The nature of concept detection makes it possible to discover related concepts through mining ground truth annotations for co-occurrences of concepts and training models. For example, the discriminative model fusion (DMF) method [21] generates a model vector based on the detection score of the individual detectors and trains a discriminative model to learn the contextual relationships, while the boosted conditional random field (BCRF) method [15] learns the contextual relationships by graph learning. In these early *offline* methods, the learning is fully supervised. Also, the learning requires explicit knowledge of the target semantics and ground truth labels to discover the contextual relationships. While this constraint is fine for concept detection where many labels are available, it is unclear how these methods can deal with the unsupervised conditions in search.

Work on using concept detectors for context fusion in search is relatively limited. Basic applications require methods of filtering out or boosting certain concept types depending on the type of query [24], [17], or investigating matches between query keywords and concept descriptions to find related concepts [7], [25]–[27]. However, deeper relationships to peripheral concepts are difficult to uncover, particularly in search, where the details of the target semantics are unknown to the system. Moreover, these relationships are often shaky, sometimes degrading search performance by uncovering relationships that are ill-suited for video retrieval [7]. As a result, it is still considered more feasible to uncover the visual co-occurrence of related concepts from the target semantics directly instead of the less meaningful lexical relationships.

B. Online Classification-Based Reranking

In view of the lack of supervision for visual search, online methods approximate the initial result of a baseline system as *pseudoground truth* to learn the contextual patterns. Rooted in pseudorelevance feedback for text search [8], [19], [28], online classification-based reranking leverages features that can potentially discriminate between higher rank and lower rank images in the initial result to determine a new ordering of the images without resorting to any offline learning or extra training data. In [5] and [6], the reranking framework is applied to low-level features such as text token frequencies, grid color moments, and image textures in a statistical manner. Authors of [7] explore the use of discriminative classifiers on a large concept lexicon [1] composed of 374 high-level concepts. The pseudopositive and pseudonegative examples are used to train a support vector machine (SVM), and the classification margin for an object is regarded as its reranked score. As shown in [7], the performance of classification-based reranking is comparable to that of supervised methods such as BCRF [15] and DMF [21]. Evaluated on TRECVID 2005, it improves the concept detection baseline from MAP 0.399 to 0.427 (a 7.0% relative improvement) and the search baseline "text-okapi" from 0.087 to 0.112 (28.7%).

However, as mentioned in Section I, classification-based reranking formulates ranking as a classification problem and thus neglects the underlying ordinal information of the ranked list. In addition, the classification approach requires an *ad hoc* mechanism to determine the threshold for noisy binary labels. To remedy such pitfalls, the proposed ordinal reranking incorporates ranking algorithms into the reranking framework to mine the co-occurrence patterns.

III. RANKING ALGORITHMS

In this section, we first introduce the *learning-to-rank* task and two famous ranking algorithms, RankSVM and ListNet, and then compare learning-to-rank and reranking.

A. Learning-to-Rank

Any system that presents ordered results to a user is performing a ranking. A common example is the ranking of search results from the search engine (e.g., Google). A ranking algorithm assigns a relevance score to each object, and ranks the object by the score. The ranking order represents the relevance of objects with respect to the query. In the literature, the task of training a ranking model which can precisely predict the relevance scores of test data is commonly referred to as learning-to-rank, which has received great interests from academia and industry [11]–[13], [31].

For learning-to-rank, a query is associated with a list of training data $D = (d_1, d_2, \ldots, d_N)$, where N denotes the size of training data and d_j denotes the jth object, and a list of manually annotated relevance scores $Y = (y_1, y_2, \ldots, y_N)$, where $y_j \in [0, 1]$ denotes the relevance score of d_j with respect to the query. Furthermore, for each object

d_j a feature vector $X_j = (X_{j1}, X_{j2}, \ldots, X_{jM})$ is extracted, where M is the dimension of the feature space. The purpose of learning-to-rank is to train a ranking model $f(\cdot)$ that accurately predicts the relevance score of test data by leveraging the co-occurrence patterns between $\mathbf{X}$ and Y. More specifically, for the training set D we obtain a list of predicted relevance score $Z = (z_1, z_2, \ldots, z_N) = (f(X_1), f(X_2), \ldots, f(X_N))$. The objective of learning is to minimize the total loss $L(Y, Z)$, where L is a loss function for ranking.

Many existing ranking algorithms take object pairs as instance in learning. These *pairwise* approaches formulate the learning task as classification of object pairs into two categories (correctly ranked and incorrectly ranked) and train classification models for ranking. The use of SVM, boosting, or neural network as the classification model leads to the methods RankSVM [11], RankBoost [29], and RankNet [30]. Though the pairwise approach offers advantages, it ignores the fact that ranking is a prediction task on a list of objects. In addition, the pairwise approach is time-consuming as the operation on every possible pair is of $O(N^2)$ complexity.

The listwise approach *ListNet* proposed in [20] conquers these shortcomings by using score lists directly as learning instances and minimizing the listwise loss between the initial list and the reranked list. In this way, the optimization is performed directly on the list, and the computational cost can be reduced to $O(N)$, making online reranking applications possible. Our experiment (described later in Section VI) shows ListNet is surprisingly efficient and even outperforms the well known pairwise approach RankSVM.

More specifically, to define a listwise loss function, authors of [20] first employs *top-one probability* to transform a list of ranking scores into a probability distribution. Given the scores of all the objects, the top-one probability $P(y_j)$ of an object d_j represents the probability of d_j being ranked on the top

$$P(y_j) = \frac{\Phi(y_j)}{\sum_{n=1}^{N} \Phi(y_n)} = \frac{\exp(y_j)}{\sum_{n=1}^{N} \exp(y_n)} \tag{1}$$

where $\Phi(\cdot)$ is an increasing and strictly positive function such as the exponential function [20]. Since the list of scores is modeled as a probabilistic distribution, a metric such as the cross entropy can be used to measure the distance (listwise loss) between the original score list and the predicted one

$$L(Y, Z) = -\sum_{j=1}^{N} P(y_j) \log(P(z_j)). \tag{2}$$

In [20], a linear neural network model is employed as the ranking model, predicting the ranking score in the form of a linear weighted sum

$$z_j = f(X_j) = \langle W, X_j \rangle \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes inner product and $W = (w_1, w_2, \ldots, w_M)$ is a weighting vector. To minimize (2), we can derive its gradient with respect to W as

$$\Delta W = \frac{\partial L(Y, Z)}{\partial W} = \sum_{j=1}^{N} \big(P(z_j) - P(y_j)\big) X_j \tag{4}$$

and then use the gradient descent procedure to update W at a learning rate η

$$W \leftarrow W - \eta \times \Delta W \tag{5}$$

where W is initially set to zero. The learning process terminates when the change in W is less than a convergent threshold δ. The values of η and δ are determined empirically, and a parameter sensitivity test over them is presented in Section VI-B.4.

B. Learning-to-Rank Versus Reranking

Reranking and learning-to-rank differ in a number of aspects. First, while learning-to-rank requires a great amount of supervision, reranking takes an unsupervised fashion and requires no ground truth from the initial concept detection or search results. The online training set D is made up of the objects of the initial results. The associated relevance scores assigned by the baseline method are taken directly as the *pseudo ground truth* Y. No manual labeling, extra training data or offline learning is needed. Second, for learning-to-rank the ranking algorithm $f(\cdot)$ is trained in advance (the training data can consist of multiple queries) to predict the relevance scores for arbitrary queries, while for reranking $f(\cdot)$ is trained at runtime by cross validation (described later) specifically for each query.

As described in Section II, existing approaches mine the co-occurrence patterns via statistical [5], [6] or classification [7], [19], [28] methods. Despite that both learning-to-rank and reranking explore ordinal information, little effort has been made to incorporate ranking algorithms into the online reranking framework. To our best knowledge, this paper represents one of such attempts.

IV. Ordinal Reranking

The input to ordinal reranking is a list of objects D and the corresponding relevance scores Y assigned by a baseline model (for either search or concept detection). We assume that feature extractions (e.g., concept detections) for each visual object are computed in advance. For these N objects in D, the corresponding M-dimensional features can form an $N \times M$ feature matrix $\mathbf{X}$. A concept score is a real value in [0, 1] that indicates the confidence of existence of the specific concept. The major steps of ordinal reranking are as follows (also illustrated in Fig. 1).

1) *Concept Selection:* Select informative concepts via a concept selection method (cf. Section V) to reduce the feature dimension to M'.
2) Randomly partition the dataset into F-folds $D = \{D^{(1)}, D^{(2)}, \ldots, D^{(F)}\}$.
3) *Employment of Ranking Algorithms:* Hold out one fold of data $D^{(1)}$ as the test set and train the ranking algorithm $f^{(i)}$ using the remaining data. Predict the new relevance scores $Z^{(i)}$ of the test set $D^{(i)}$. Repeat until each fold is held out for testing once. The predicted scores of different folds of objects are then combined to form the new list of scores $Z = \{Z^{(1)}, Z^{(2)}), \ldots, Z^{(F)}\}$.

4) *Rank Aggregation:* After normalization, the initial score y_j and new score z_j (for object d_j) are fused to produce a merged score s_j by taking a weighted average as follows:

$$s_j = (1 - \alpha) y_j + \alpha z_j \tag{6}$$

where $\alpha \in [0, 1]$ denotes a fusion weight on the initial and reranked scores; $\alpha = 1$ means totally reranked.

5) *Rerank:* Sort the fused scores to output a new ranked list for the target semantics.

In this paper, we focus on leveraging hundreds of pre-trained concept detectors for context fusion. Therefore, we use concept scores to form the feature space and apply concept selection methods to select informative concepts. In practice, thanks to the generality of the ordinal reranking approach, we can also utilize contextual cues other than high-level concept scores such as low-level visual features, time stamps, geo-tags, etc. For example, in [2] we use visual words [35] and geo-tags as the contextual cues for reranking the result of a consumer photo search system.

To accommodate the supervised ranking algorithms to the unsupervised environment of reranking, we employ the F-fold cross validation technique [7] to partition the dataset, and train F ranking algorithms with different folds of data held out as the test set. Though the new scores Z are not assigned by a unified ranking algorithm, the nature of cross validation ensures that F–2 folds of data are commonly used in the training of two different ranking algorithms.

A ranking algorithm predicts the relevance scores of objects by exploiting the co-occurrence patterns learnt from the training data. We then use (6) for rank aggregation of the initial relevance scores and the newly predicted scores. Such a linear fusion model, though simple, has been shown adequate to fuse visual and text modalities in video retrieval and concept detection [17], [18]. The fusion weight α controls the degree of context fusion and may be influential to the reranked result. In our experiments described in Section VI.*B.4*, variant fusion weights are tested experimentally.

V. Concept Selection

We further investigate concept selection methods to remove irrelevant or redundant features and thus enhance accuracy. We improve the concept tf-idf (c-tf-idf) proposed in [10] by incorporating the ordinal information of the initial list and emphasizing the contribution of higher rank objects. We first briefly review the existing feature selection methods and then describe the proposed *weighted c-tf-idf* (wc-tf-idf).

As reported in [10] and [7], the performance of a video search model can degenerate significantly as the feature dimension increases arbitrarily. However, the feature selection methods used in most existing ranking models are originally designed for classification rather than for ranking. Applying these selection methods may be problematic due to the distinct problem definition and objective function between classification and ranking. There has been rare method of feature selection specifically proposed for ranking, leaving the work on feature selection for ranking a still unsolved problem [31]. One exception is proposed in [32], where feature selection is formulated as an optimization problem and the evaluation measures in ranking are utilized to measure feature importance. However, the feature selection method needs ground truth data and thus is inapplicable to online reranking.

A promising solution is the totally unsupervised c-tf-idf measurement proposed in [10]. As the name implies, c-tf-idf is adapted from the best known term-informativeness measurement tf-idf [14]. Viewing visual objects as documents and concept scores as visual term frequencies, we construct a document–concept occurrence table from a list of visual objects and the associated concept scores, and define the c-tf-idf of concept c in a query q as follows:

$$\text{c-tf-idf}(c, q) = freq(c, q) \log\left(\frac{T}{freq(c)}\right), \quad c \in C \tag{7}$$

where $freq(c, q) = \sum_{j=1}^{N} X_{jc}$ is the occurrence frequency of c in the initial reranked list, X_{jc} is the concept score of c in d_j (estimated by a pre-trained concept detector), N is the length of the initial list, and C denotes the concept set. The denominator $freq(c) = \sum_{j=1}^{T} X_{jc}$ is the occurrence frequency of c in the whole corpus. T is the size of corpus, and typically $T \gg N$ because the initial list is a subset of objects that are considered relevant. The intuition of (7) is that the relevance of a concept increases proportionally to the frequency it appears in the return list of a query, but is offset by the frequency of the concept in the entire corpus to filter out common concepts (such as *face* and *indoor*). In this way, c-tf-idf offers a good combination between popularity (idf) and specificity (tf) [14].

However, because c-tf-idf considers each object in the initial list as equally relevant to the target semantics, the underlying ordinal information is totally neglected. Since the higher rank objects are more relevant to the target semantics, they should be weighted more than the lower rank ones. To this end, we propose to utilize the initial relevance scores to weight the concept score of each object as follows:

$$\text{wc-tf-idf}(c, q) = \sum_{j=1}^{N} y_j X_{jc} \log\left(\frac{T}{\sum_{j=1}^{T} X_{jc}}\right). \tag{8}$$

After sorting wc-tf-idf(c, q) in descending order, the top ranked concepts are selected to train the ranking algorithms. Note that wc-tf-idf is generic and applicable to other contextual cues such as low-level visual features.

VI. Experimental Results

A. Experiment

A series of experiments are conducted on the TRECVID 2005 (TV05) dataset to give a comprehensive evaluation of the proposed reranking framework.[4] The TV05 dataset

[4] 'TRECVID' stands for TREC Video Retrieval Evaluation [1], the goal of which is to promote content-based video analysis and retrieval via open, metrics-based evaluation. The TRECVID dataset has been widely used as a benchmark for evaluating video search and concept detection methods. We use TRECVID 2005 here so that we can compare our method with others [6], [7].

TABLE II
PERFORMANCE COMPARISON OF VARIOUS RERANKING METHODS ON THE TRECVID 2005 SEARCH TASK USING ''TEXT-OKAPI'' AS THE BASELINE

#	Reranking Algorithm	Feature Set	Feature Selection	MAP	Improvement (%)	Time/Query
1	Baseline	Text-only	–	0.087	–	–
2	IB [6]	Low-level	–	0.105	20.7	18 s
3	SVM [7]	cp374	Mutual information	0.112	28.7	17 s
4	RankSVM	cp39	–	0.103	18.4	1 h
5	ListNet	cp39	–	0.113	30.0	0.2 s
6	ListNet	low-level	–	0.105	20.7	0.9 s
7	ListNet	cp374	–	0.116	33.3	1.4 s
8	ListNet	cp374	c-tf-idf	0.118	35.6	0.4 s
9	ListNet	cp374	wc-tf-idf	0.121	40.0	0.4 s

consists of 277 international broadcast news video programs and accumulates 170 h of videos from six channels in three languages (Arabic, English, and Chinese). The time span is from October 30 to December 1, 2004. The automatic speech recognition and machine translation transcripts are provided by the National Institute of Standards and Technology. The video data is segmented into shots and each shot is represented by a few keyframes (subshots). In the following experiments, we evaluate the performance at shot level in terms of average precision (AP), which approximates the area under a non-interpolated recall/precision curve. Since AP only shows the performance of a query, we use MAP, which is simply the mean of APs for multiple queries, to measure average performance over sets of different queries in the test data. See more explanations in [1].

We first apply the reranking approach to the search task, where 24 query topics are provided with ground truth annotations. Since TRECVID evaluates the search results over the top 1000 shots, we use the top 1300 subshots (which typically encompass the top 1000 shots) returned by the text-based search method "text-okapi" [6] for reranking.

We also apply ordinal reranking for the concept detection of the 39 LSCOM-Lite concepts [4] over a set sampled from the TV05 development data.[5] Since TRECVID evaluates the high-level concept detection results over the top 2000 shots, we use the top 2600 subshots returned by the baseline detection method from [17] for reranking. In a supervised fashion, the concept detection accuracy is generally much higher than that from the search baseline.

For feature representation, we adopt the detection scores of pre-trained concept detectors [4] for the LSCOM (cp374) and LSCOM-Lite (cp39) lexicons [17] to provide high-level semantics. The LSCOM concept lexicon is a set of 374 visual concepts which were annotated over an 80-h subset of the TRECVID data. The LSCOM-Lite lexicon is with 39 concepts and is an early version of LSCOM. Low-level visual features (low-level) including 5×5 grid color moments and 4×6 Gabor textures [6] are also included to compare against the high-level concept scores. Though they are primitive feature representations, prior work such as [6], [17] has shown their excellence in image retrieval and concept detection.

[5]Dataset available online: http://mpac.ee.ntu.edu.tw/~yihsuan/reranking/.

The implementation of RankSVM is based on the software SVM$^{\text{light}}$ [33] with default parameters. ListNet is implemented in MATLAB. The programs are executed on a regular Intel Pentium server.

B. Reranking for Video Search

1) *Comparison of Variant Reranking Methods and Feature Sets:* We first conduct experiments on TV05 video search task with variant reranking methods and feature sets without feature selection. Empirically, we set the fusion weight α to 0.5 for simplicity and use fivefold cross validation to conduct reranking. The learning rate η and convergent threshold δ for ListNet are empirically set to 0.005 and 1e$-$4, respectively (parameter sensitivity tests are presented later). While the MAP of the text-based search baseline "text-okapi" is 0.087, existing methods [6], [7] improve the MAP to 0.105 and 0.112 respectively, as shown in the second and third rows of Table II. Note that [6] uses low-level visual features without feature selection, whereas [7] utilizes mutual information to select the 75 most informative concepts among cp374 for reranking.

We first compare the performance of ordinal reranking with different ranking algorithms, namely, RankSVM and ListNet. As shown in the fourth row of Table II, RankSVM is extremely time-consuming and thus get abandoned in the following experiments. On the contrary, thanks to the linear kernel, ListNet is surprisingly efficient and takes less than one second to rerank a single query. The efficiency of ListNet makes it superior to the method described in [6], which needs a clustering process, and to the method described in [7], which uses non-linear optimization and *ad hoc* thresholding. Moreover, the fact that ListNet improves the MAP to 0.113 with a small concept lexicon cp39 further demonstrates its effectiveness in reranking.

We then evaluate the performance of ListNet with variant feature sets (rows 5–7 of Table II). Despite the less salient performance gains that can be provided, low-level visual features still offer contextual cues that augment the baseline text-based methods and improve the MAP to 0.105. It is not surprising that reranking based on low-level visual features does not perform as well as that based on high-level concepts since concepts tend to capture both the visual similarities and the semantic correlations. In addition, the visual patterns of target semantics may not be consistent or evident enough, bringing noises to the reranking procedure. For example, a

TABLE III
TOP POSITIVELY WEIGHTED CONCEPTS OF LISTNET AND TOP SELECTED CONCEPTS OF WC-TF-IDF AMONG LSCOM 374 CONCEPTS FOR VIDEO SEARCH

Query	Baseline AP	ListNet Without Feature Selection		ListNet With Top 150 Ranked Concepts Selected by wc-tf-idf	
		AP	Top Positively Weighted Concepts	AP	Top Selected Concepts
condoleeza_rice (149)	0.098	0.085	Military building, U.S. flags	0.089	Crowd, non-uniformed fighters, parade, people marching
omar_karami (151)	0.443	0.532	Interview, government leader	0.571	Singing, meeting, conference room, furniture, interview
hu_jintao (152)	0.211	0.188	Flag-U.S., network logo, horse	0.185	Asian people, Hu jintao, Colin Powell, government leader
tony_blair (153)	0.357	0.554	Hu Jintao, judge, Colin Powell	0.588	Person, room, furniture, actor, sketches, civilian people
mahmoud_abbas (154)	0.259	0.391	Ground crew, smoke stack	0.428	Singing, ground crew, baby, outer space, conference room
tennis_court (156)	0.021	0.027	Tennis, text labeling people	0.030	Athletes, studio with anchorperson, Caucasians, basketball
boat (164)	0.169	0.233	Cigar boats, river, lakes	0.235	Lakes, airplane flying, exploding ordinance, waterscape
soccer_goal (171)	0.263	0.407	Soccer, lawn, stadium	0.480	Athletes, basketball, indoor sports venue, stadium, running

named person may wear clothes of different colors and appear in diverse locations.

Ordinal reranking with the semantic-richer cp374 feature set, though improves the MAP further to 0.116, does not yield significant performance gain against that of cp39. It might suggest the necessity of concept selection. Because ListNet forms the prediction result by the weighted sum of concept scores, the weight assigned to each concept can be regarded as the relevance to the query settled by ListNet. Positive (negative) weight implies positive (negative) correlation. Therefore, we examine the weights to evaluate whether ListNet discovers perceptually related concepts. Table III (the left and middle parts) lists the top positively weighted concepts for each query and the resulting APs before and after reranking. Some highly weighted concepts are indeed relevant to the queries, e.g., the concepts *soccer*, *lawn*, and *stadium* are brought out for the query "soccer_goal (171)." However, some counterexamples do exist, such as *flag − U.S.*, *network logo*, and *horse* for "hu_jintao (152)." These erroneously weighted concepts offer little contextual information for reranking, and even lead to serious performance degradation.

2) *Performance of Concept Selection:* To select informative concepts for each query, we apply c-tf-idf and wc-tf-idf to rank cp374 according to the relevancy to the target query, and evaluate the resulting MAP by varying numbers of concepts for reranking. It can be observed from Fig. 2 that reranking based on the concepts selected by wc-tf-idf can achieve higher MAP with fewer concepts than those selected by c-tf-idf, showing that wc-tf-idf is more adept at selecting informative concepts than c-tf-idf. As the number of selected concepts grows, the MAPs first reach a plateau since most informative concepts have been selected, and then begin to degrade after too much irrelevant concepts are added. It is also observed that, with the 150 most informative concepts selected by wc-tf-idf, ListNet improves the MAP up to 0.121, which

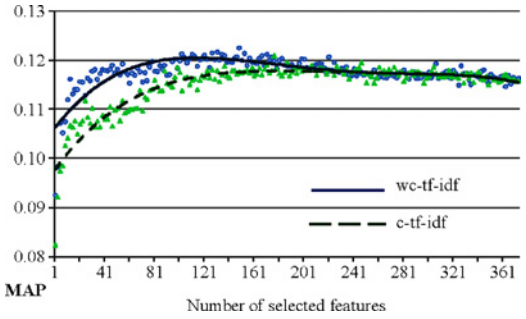

Fig. 2. Reranking result of the TRECVID 2005 search baseline "text-okapi" using ListNet with different numbers of top ranked concepts selected by wc-tf-idf and c-tf-idf. It can be observed that wc-tf-idf achieves higher MAP with fewer concepts, showing its superior concept selection ability.

outperforms existing methods significantly. As we summarized in the ninth row of Table II, the incorporation of concept selection procedure does not add much computational burden, and the reranking of a single query can still be finished in roughly 0.4 s.

In the right part of Table III shows the top ranked concepts among cp374 selected by wc-tf-idf and the resulting APs by using the 150 most informative concepts for ordinal reranking. Obviously, many of the concepts judged relevant to a specific query do make sense. For example, the concept *athlete* is ranked high for both "soccer_goal (171)" and "tennis_court (156)," despite the low AP of the latter query. *Conference room* and *government leader* are ranked high for the name persons, which are most politicians. However, there are still some counterexamples, e.g., *singing* for "mahmoud_abbas (154)" and *non-uniformed fighters* for "condoleeza_rice (149)." We believe this mismatch is reasonable since the accuracies of the LSCOM concept detectors are not equally well, and can bring noises to the concept scores to which wc-tf-idf is applied. Noises may also come from the incorrect ordinal information

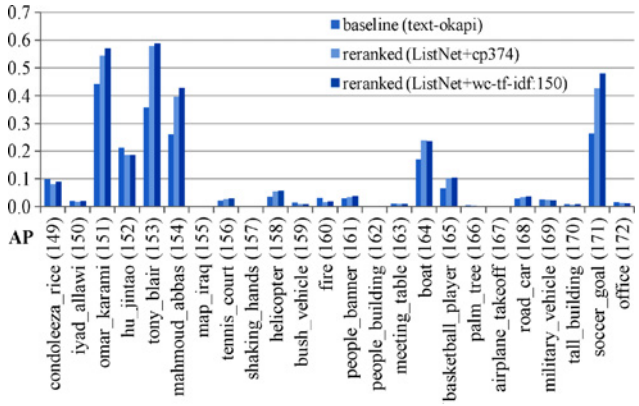

Fig. 3. Average precisions of baseline and reranked search results for each query in TRECVID 2005. The reranking algorithm is ListNet, with 150 most informative concepts selected from LSCOM 374 concepts by wc-tf-idf. Ordinal reranking improves the result of almost every query and improves the MAP from 0.087 to 0.122 (40% relative improvement).

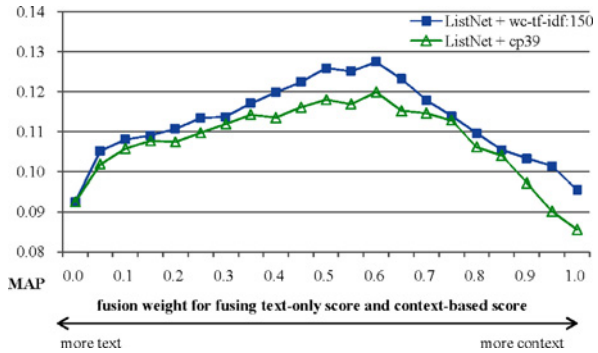

Fig. 4. MAPs for applying ListNet to rerank search baseline "text-okapi" with two feature sets as we change the fusion weight α from 0 to 1 at an increasing step of 0.05. Optimal performance is achieved by setting α at 0.6.

set by the baseline model. From Table III it can also be observed that, for the query "hu_jintao (152)," despite the top selected concepts are mostly correct, the AP is not improved after feature selection. The reason is we have fixed the number of selected concepts to be 150 for all the queries. Actually, we found the overall performance can be further improved if the number of selected concepts is query-dependent. For example, if only the top three ranked concepts are used for reranking "hu_jintao (152)," the AP can reach 0.192. Yet we leave it as part of the future works.

3) *Discussion of the Reranking Performance for Video Search:* Fig. 3 depicts the APs achieved by the baseline method, reranking using ListNet with cp374 as the feature set, and reranking using ListNet with the 150 most informative concepts selected by wc-tf-idf. Remarkably, the performance improvements of ordinal reranking over the baseline are consistent—almost all queries are improved. Among them, salient improvements are observed for queries with higher initial AP, such as "omar_karami (151)," "tony_blair (153)," and "soccer_goal (171)." Concept selection by wc-tf-idf further enhances the result, especially for "soccer_goal (171),"which has strong contextual links with many concepts.

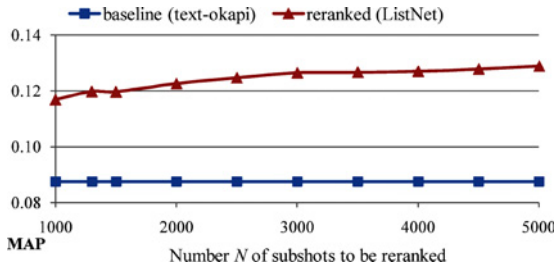

Fig. 5. MAPs for applying ListNet with different numbers N of subshots to be reranked. This result shows the MAP, which is evaluated using the top 1000 subshots, can be further improved by using a larger N, which provides more training data and increase the possibility to improve the recall rate by retrieving subshots which are ranked lower in the initial list.

However, as existing reranking method [7], for queries which lack viable methods for getting a good initial search result, such as "map_iraq (155)" and "tall_building (170)," ordinal reranking does not offer improvements since the contextual patterns are difficult to discover. This may be an inherent limitation of the reranking methodology, which heavily relies on the quality of the initial model. We do not discuss this issue further, leaving it as part of the future research.

4) *Parameter Sensibility:* We are also interested in the performance with different values of weight α for fusing text-based relevance scores Y and context-based reranked scores Z in (6). To analyze the impact of fusion weights, we compare the performance with different fusion weights ranging from 0.0 (totally text-based) to 1.0 (totally reranked) with an increasing step of 0.05 and plot the results in Fig. 4. Interestingly, we discover the influence of α similar to what has been reported in [5]. When α is close to 1, the reranking process relies almost entirely on the contextual similarities and ignores the text search prior; hence, the performance degrades sharply. When $\alpha = 1$ (totally reranked), the performance of the process is similar to that of the purely text-based method. In addition, the text modality and context modality carry important information, and the fusion of both gives rise to optimal reranking, showing the two modalities are quite complementary. The same phenomenon is observed when either cp39 or cp374 is used as the feature set.

We then conduct parameter sensibility test on the values of the learning rate η and the convergent threshold δ for ListNet. Being used in (5), η controls the degree of updating of the weights W and influences the convergence time. As shown in Table IV, the performance of ListNet is rather invariant to the changes of η as long as it is set to a moderately small value. A large η may result in an oscillation of W and degrade the performance. On the other hand, δ controls the degree of the ranking algorithm fitting the training data and also influences the convergence time. It can be observed in Table IV that setting δ too small overfits the data and has negative effect on the reranking performance. To balance the convergence time and the reranking performance, we have set η and δ to 0.005 and 1e−4, respectively (the cell marked with * in Table IV).

Finally, as N determines the number of subshots to be reranked and the size of training data, it is also interesting to know whether we can improve the MAP by using a large

TABLE IV
RERANKING PERFORMANCE OF LISTNET WITH VARIANT PARAMETER SETTINGS FOR THE TRECVID 2005 SEARCH TASK

Convergent Threshold δ	Learning Rate η				
	1e−4	1e−3	5e−3	1e−2	1e−1
1e−3	0.121	0.121	0.120	0.120	0.092
1e−4	0.121	0.121	0.121*	0.120	0.085
1e−5	0.119	0.119	0.115	0.115	0.082
1e−6	0.120	0.114	0.111	0.110	0.081

*The parameter settings used in other experiments in this paper.

TABLE V
RERANKING PERFORMANCE OF LISTNET WITH VARIANT FEATURE SETS FOR TRECVID 2005 CONCEPT DETECTION OVER THE DETECTION BASELINE

Method	Feature Set	Feature Selection	Number of Feature	MAP	Improvement (%)	Time/Concept
Baseline	–	–	–	0.369	–	–
ListNet	cp39	–	39	0.379	2.7	0.7s
	cp374	–	374	0.410	11.1	3.7s
	cp374	wc-tf-idf	150	0.406	10.0	1.4s
	cp374	wc-tf-idf	25	0.413	11.9	0.4s

TABLE VI
TOP SELECTED CONCEPTS BY WC-TF-IDF AMONG LSCOM 374 CONCEPTS FOR CONCEPT DETECTION

Concept	Top Selected Concepts	Concept	Top Selected Concepts
Walking_running	Walking running, crowd, basketball, walking	Crowd	Crowd, people marching, funeral, parade, cheering
Explosion_fire	Exploding ordinance, explosion fire, street battle	Court	Sketches, Colin Powell, court, interview on location
Maps	Maps, studio, news studio, studio with anchorperson	Desert	Desert, street battle, rocky ground, weapons
Flag-U.S.	Studio with anchorperson, flag-U.S., Colin Powell	Entertainment	Singing, entertainment, celebrity entertainment, room
Building	Building, road, office building, urban scenes	Meeting	Meeting, furniture, conference room, Colin Powell
Waterscape_waterfront	Oceans, waterscape waterfront, waterways, lakes, logo full screen, commercial advertisement, boat ship	Corporate_leader	Corporate leader, interview on screen, person, interview sequences, Colin Powell, government leader
Mountain	Mountain, valleys, hill, rocky ground, landscape	Face	Person, studio with anchorperson, civilian person
Prisoner	Sketches, prisoner, election campaign greeting, person	Military	military personnel, soldiers, street battle, military
Sports	Athlete, basketball, sports, indoor sport venue, baseball	Natural_disaster	natural disaster, rocky ground, still image, flood
Car	Ground vehicle, road, car, vehicle, streets	Truck	Road, ground vehicle, truck, vehicle, daytime outdoor
Charts	Charts, text on artificial background, logo full screen	Vegetation	Trees, vegetation, daytime outdoor, lawn, golf course
Computer_tv_screen	Studio with anchorperson, studio, news studio, computer tv screen, female anchor, male anchor	Weather	Weather, maps, studio, news studio, charts, text on artificial background, studio with anchor person
Animal	Animal, birds, valleys, oceans, logos full screen, hill	Outdoor	Daytime outdoor, outdoor, road, military personnel

N (which is previously set to 1300). We thus vary N from 1000 to 5000 at a step size of 500 and evaluate the MAP over the top 1000 subshots of the reranked result. Result is shown in Fig. 5, which indicates the MAP is indeed improved with larger N; the MAP is improved to 0.129 (47% relative improvement to the text baseline) when setting N to 5000. Using large N provides more training data to the learning-to-rank algorithms, and increases the possibility to improve the recall rate by retrieving subshots which are ranked lower in the initial list. It may be also possible to further improve the recall rate by applying the learnt contextual patterns to the whole corpus to retrieve subshots that are not included in the initial list.

C. Reranking for Concept Fusion

We also study the performance of using context of 374 concept detectors to improve the detection accuracy of each of the 39 LSCOM-Lite concepts. The use of cp374 for context fusion of the detection of cp39 is reasonable since cp374 covers richer high-level semantics. Note that the concept detection methods utilized in cp374 [4] and cp39 [17] are different since the latter specifically included some parts-based detection models. Thus the detection results for the same concept can be different. The same parameter setting for video search is used here. As [7] reported, existing reranking methods, including the offline ones and the classification-based ones, offer 5–7% performance gains, which are less than those

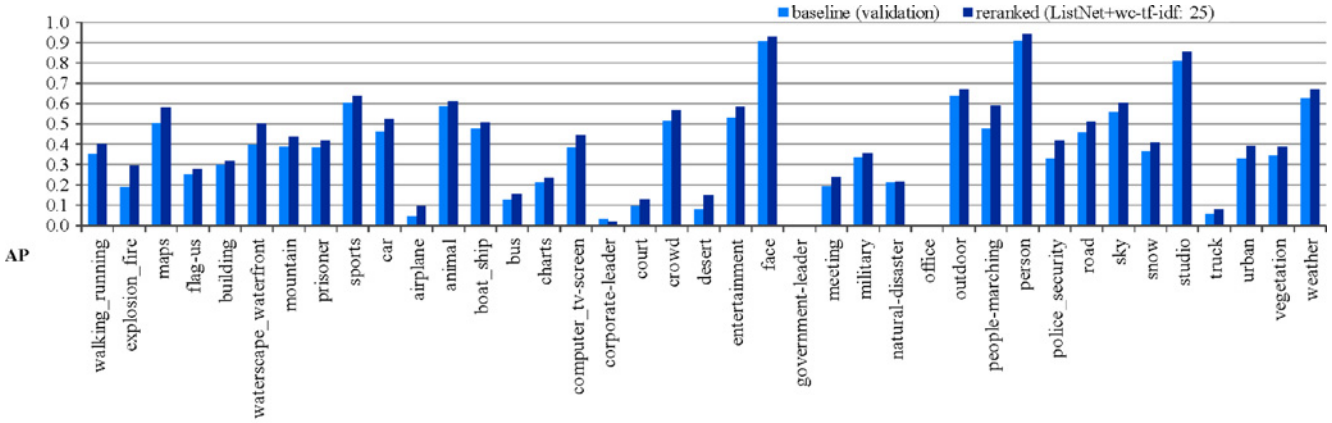

Fig. 6. Average precision of baseline and reranked results for each target concept in the TRECVID 2005 benchmark. Here ListNet is adopted as the reranking algorithm, with 25 most informative concepts selected from LSCOM 374 concepts by wc-tf-idf. Ordinal reranking improves the result of almost every concept detector and improves the MAP from 0.369 to 0.413 (12% relative improvement).

for visual search since the initial concept detection accuracy has been relatively high (e.g., MAP 0.369 in Table V).

Results shown in Table V indicate that the reranking approach is also effective for refining the baseline concept detectors: using cp374 as the feature set, ListNet improves the MAP from 0.369 to 0.410 (11.1%) over the detection baseline. If the 25 most informative concepts selected by wc-tf-idf are used, the MAP is further improved to 0.413 (11.9%), which significantly outperforms existing context fusion methods. In addition, ordinal reranking is still remarkable efficient for concept fusion, taking less than one second to refine the result of a concept. Moreover, as shown in Fig. 6, the performance improvements of ordinal reranking over the baseline are also consistent for concept fusion.

From Table V it can also be observed that context fusion based on cp39 only slightly improves the result. This is not surprising since cp39 is actually an essential subset of cp374 and thus the contextual links between the cp39 concepts are much limited. The other interesting observation is related to the optimal number of concepts for use in ordinal reranking. While the optimal number of concepts is around 150 for video search, the optimal number of concepts is merely 25, which implies that most informative concepts are successfully selected by wc-tf-idf. The cause of this notable effectiveness is twofold. First, the contextual patterns among cp374 (LSCOM concepts) and the target semantics (39 LSCOM-Lite concepts) are in nature strong. Second, the initial accuracies of the concept detectors are already high, making the contextual patterns easy to be discovered.

We also tabulate the top ranked concepts among cp374 in Table VI, without excluding the identical concept as the target concept in cp374. As Table VI shows, most target concepts rank the identical one in top three, and the contextual relationship among top selected concepts are intuitively correct. For example, *people marching*, *funeral*, and *parade* are ranked high for "Crowd," and *ground vehicle*, *road*, *streets* are ranked high for "Car." In addition, we observe that wc-tf-idf is also adept at discovering contextual links which query expansion or keyword matching easily fail to. For example, *street battle* and *weapons* are ranked high for both "Desert" and "Explosion_fire," while *map*, *charts*, and *studio* are ranked high for "Weather." These relationships may be less salient literally, yet are essentially correct since the makeup of TRECVID videos are mostly news videos [7].

VII. Conclusion

In this paper, we have exploited the contextual information for visual search and concept detection and proposed a novel reranking algorithm called ordinal reranking for mining the co-occurrence patterns between the target semantics and the extracted features. This ranking-based reranking algorithm is more effective and efficient than existing reranking methods. Moreover, because ordinal reranking directly optimizes the ordering of an initial list obtained by a baseline system, it is free of *ad hoc* thresholding for noisy binary labels and requires no extra offline learning processes or training data. Besides, as there has been rare feature selection measure specifically designed for reranking, we also propose a novel measurement, wc-tf-idf, to select informative concepts and further improve the performance of reranking.

Because ordinal reranking is largely unsupervised, it can be applied equally well to context fusion in both concept detection and video search tasks. An extensive performance study is conducted on the TRECVID 2005 benchmark to evaluate the performance of ordinal reranking and concept selection for the two tasks. Results show that ordinal reranking is much more efficient and effective than existing reranking methods and improves the MAP up to 40% over the text-based search results and 12% over the concept detection baselines.

The proposed ordinal reranking approach is general enough to be applied to problems in other multimedia domains such as media-rich social networks and blogs that have strong contextual links between pieces of information that come from multiple sources.

References

[1] National Institute of Standards and Technology. *Text Retrieval Conference Video Retrieval Evaluation* [Online]. Available: http://www-nlpir.nist.gov/projects/trecvid

[2] Y.-H. Yang, P.-T. Wu, C.-W. Lee, K.-H. Lin, W. H. Hsu, and H.-H. Chen, "ContextSeer: Context search and recommendation at query time for shared consumer photos," in *Proc. Assoc. Comput. Mach. Multimedia*, 2008, pp. 199–208.
[3] M. Naphade *et al.*, "Large-scale concept ontology for multimedia," *IEEE Multimedia Mag.*, vol. 13, no. 3, pp. 86–91, Jul.–Sep. 2006.
[4] A. Yanagawa *et al.*, "Columbia University's baseline detectors for 374 LSCOM semantic visual concepts," Columbia Univ., New York, NY, ADVENT Tech. Rep. #222-2006-8, 2007.
[5] W. H. Hsu, L. Kennedy, and S.-F. Chang "Video search reranking through random walk over document-level context graph," in *Proc. Assoc. Comput. Mach. Multimedia*, 2007, pp. 971–980.
[6] W. H. Hsu, L. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. Assoc. Comput. Mach. Multimedia*, 2006, pp. 35–44.
[7] L. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," in *Proc. Assoc. Comput. Mach. Int. Conf. Image Video Retrieval*, 2007, pp. 333–340.
[8] R. Yan, R. Jing, and A. Hauptmann, "Multimedia search with pseudorelevance feedback," in *Proc. Assoc. Comput. Mach. Int. Conf. Image Video Retrieval*, 2003, pp. 238–247.
[9] J. Battelle, *The Search: How Google and Its Rivals Rewrote the Rules of Business and Transformed Our Culture*. New York: Portfolio Trade, 2005.
[10] X. Li, D. Wang, J. Li, and B. Zhang, "Video search in concept subspace: A text like paradigm," in *Proc. Assoc. Comput. Mach. Int. Conf. Image Video Retrieval*, 2007, pp. 603–610.
[11] R. Herbrich, T. Graepel, and K. Obermayer, "Support vector learning for ordinal regression," in *Proc. Int. Conf. Artif. Neural Netw.*, 1999, pp. 97–102.
[12] T. Joachims, "Optimizing search engines using clickthrough data," in *Proc. Assoc. Comput. Mach. Special Interest Group Knowl. Discovery Data Mining*, 2002, pp. 133–142.
[13] Y. Cao, J. Xu, T.-Y. Liu, H. Li, Y. Huang, and H.-W. Hon, "Adapting ranking SVM to document retrieval," in *Proc. Assoc. Comput. Mach. Special Interest Group Inform. Retrieval*, 2006, pp. 186–193.
[14] A. Aizawa, "An information-theoretic perspective of tf-idf measures," *Inform. Process. Manage.*, vol. 39, pp. 45–65, 2003.
[15] W. Jiang, S.-F. Chang, and A. C. Loui, "Context-based concept fusion with boosted conditional random fields," in *Proc. IEEE Int. Conf. Acoustics Speech Signal Process.*, vol. 1. 2007, pp. 949–952.
[16] C. G. Snoek *et al.*, "The MediaMill TRECVID 2006 semantic video search engine," in *Proc. Natl. Inst. Standards Technol. Text REtrieval Conf. Video Retrieval Eval. Workshop*, 2006.
[17] S.-F. Chang *et al.*, "Columbia University TRECVID 2005 video search and high-level feature extraction," in *Proc. Natl. Inst. Standards Technol. Text REtrieval Conf. Video Retrieval Eval. Workshop*, 2005.
[18] M. Campbell *et al.*, "IBM Research TRECVID 2006 video retrieval system," in *Proc. Natl. Inst. Standards Technol. Text REtrieval Conf. Video Retrieval Eval. Workshop*, 2006.
[19] T.-S. Chua *et al.*, "TRECVID 2004 search and feature extraction task by NUS PRIS," in *Proc. Natl. Inst. Standards Technol. Text REtrieval Conf. Video Retrieval Eval. Workshop*, 2004.
[20] Z. Cao, T. Qin, T.-Y. Liu, M.-F. Tsai, and H. Li, "Learning to rank: From pairwise approach to listwise approach," in *Proc. IEEE Int. Conf. Machine Learning*, 2007, pp. 129–136.
[21] J. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *Proc. IEEE Int. Conf. Multimedia Expo*, vol. 2. Jul. 2003, pp. 445–448.
[22] C. Snoek, M. Worring, D. Koelma, and A. Smeulders, "Learned lexicon-driven interactive video retrieval," in *Proc. Assoc. Comput. Mach. Int. Conf. Image Video Retrieval*, 2006, pp. 11–20.
[23] W. Jiang, S.-F. Chang, and A. C. Loui, "Active context-based concept fusion with partial user labels," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2006, pp. 2917–2920.
[24] A. G. Hauptmann *et al.*, "Multilingual broadcast news retrieval," in *Proc. Natl. Inst. Standards Technol. Text REtrieval Conf. Video Retrieval Eval. Workshop*, 2006, pp. 1–12.
[25] A. Haubold, A. Natsev, and M. Naphade, "Semantic multimedia retrieval using lexical query expansion and model-based reranking," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 1761–1764.
[26] S.-Y. Neo, J. Zhao, M.-Y. Kan, and T.-S. Chua, "Video retrieval using high-level features: Exploiting query matching and confidence-based weighting," in *Proc. Assoc. Comput. Mach. Int. Conf. Image Video Retrieval*, 2006, pp. 143–152.
[27] A. Natsev, A. Haubold, J. Tesic, L. Xie, and R. Yan, "Semantic concept-based query expansion and reranking for multimedia retrieval," in *Proc. Assoc. Comput. Mach. Multimedia*, 2007, pp. 991–1000.
[28] Y. Yang, J. Carbonell, R. D. Brown, and R. E. Frederking, "Translingual information retrieval: A comparative evaluation," in *Proc. Int. Joint Conf. Artif. Intell.*, 1997, pp. 708–715.
[29] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer, "An efficient boosting algorithm for combining preferences," in *Proc. IEEE Int. Conf. Machine Learning*, 1998, pp. 170–178.
[30] C. Burges, T. Shaked, E. Renshaw, A. Laizer, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *Proc. IEEE Int. Conf. Machine Learning*, 2005, pp. 89–96.
[31] T.-Y. Liu, J. Xu, T. Qin, W. Xing, and H. Li, "LETOR: Benchmark dataset for research on learning to rank for information retrieval," in *Proc. Assoc. Comput. Mach. Special Interest Group Inform. Retrieval Workshop Learning Rank Inform. Retrieval*, 2007, pp. 3–10.
[32] X. Geng, T.-Y. Liu, T. Qin, and H. Li, "Feature selection for ranking," in *Proc. Assoc. Comput. Mach. Special Interest Group Inform. Retrieval*, 2007, pp. 407–414.
[33] T. Joachims, "Making large-scale SVM learning practical," *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 169—184.
[34] A. K. Dey, "Understanding and using context," *Personal Ubiquitous Comput.*, vol. 5, no. 1, pp. 4–7, 2001.
[35] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. IEEE Int. Conf. Comput. Vision*, vol. 2. Oct. 2003, pp. 1470–1477.

Yi-Hsuan Yang received the B.S. degree in electrical engineering from National Taiwan University, Taipei, Taiwan, in 2006. He is currently working toward the Ph.D. degree from the Graduate Institute of Communication Engineering, National Taiwan University.

His research interests include multimedia information retrieval and analysis, machine learning, and affective computing. He has published over 20 technical papers in the above areas.

Mr. Yang won the Microsoft Research Asia Fellowship in 2008–2009.

Winston H. Hsu received the Ph.D. degree from the Department of Electrical Engineering, Columbia University, New York, NY.

He is an Assistant Professor in the Graduate Institute of Networking and Multimedia, National Taiwan University, Taipei, Taiwan, since February 2007. Prior to this, he was in the multimedia software industry for years. He is also with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan. His research interests include multimedia content analysis, image/video indexing and retrieval, machine learning and mining over large-scale databases.

Homer H. Chen (S'83–M'86–SM'01–F'03) received the Ph.D. degree in electrical and computer engineering from University of Illinois, Urbana-Champaign.

He has been with the College of Electrical Engineering and Computer Science, National Taiwan University, Taipei, Taiwan, since August 2003, as the Irving T. Ho Chair Professor. Prior to this, he held various Research and Development Management and Engineering positions with U.S. companies over a period of 17 years, including AT&T Bell Labs, Rockwell Science Center, iVast, and Digital Island. He was a U.S. Delegate for ISO and ITU Standards Committees and contributed to the development of many new interactive multimedia technologies that are now part of the MPEG-4 and JPEG-2000 standards. His professional interests include the broad area of multimedia signal processing and communications.

Dr. Chen is an Associate Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY. He served as Associate Editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, from 1992 to 1994, Guest Editor of IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, in 1999, and Associate Editor of *Pattern Recognition*, from 1989 to 1999.