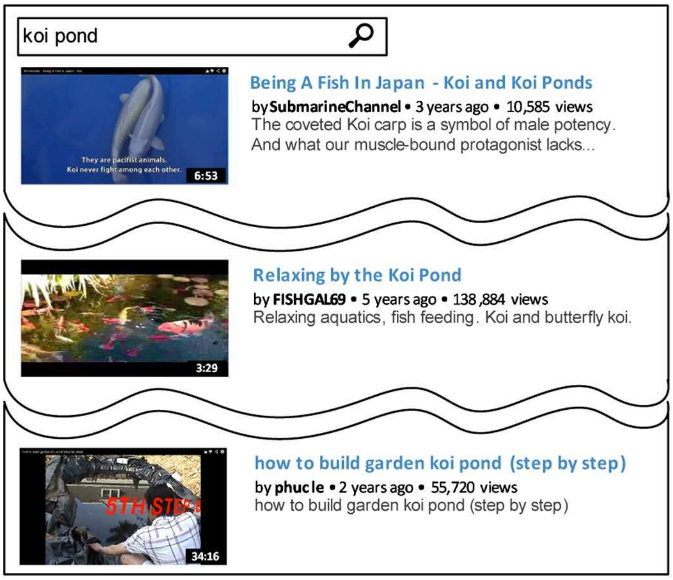
Intent-Aware Video Search Result Optimization

Christoph Kofler, Martha Larson, *Member, IEEE*, and Alan Hanjalic, *Senior Member, IEEE*

***Abstract*—Video search engines are relatively successful at returning search results that users find to be on topic. These results do not, however, completely satisfy the user's information need unless they also fulfill the user's intent, i.e., the immediate goal a user seeks to accomplish with video search. Satisfying a user's information need to its full extent poses a particular challenge to video search engines because user intent is often not explicitly reflected in the query. In this paper, we propose a multimodal approach that addresses this challenge by refining the results lists returned by a mainstream video search engine in order to optimally capture user intent. Our approach is based on the insight that the results lists returned by video search engines do contain videos that satisfy user's intent, but that videos with the highest potential for satisfaction are often buried within or scattered over the results list. The proposed approach consists of three steps. In the first step, it analyzes the initial results list to determine the intent distribution pattern. On the basis of this pattern, in the second step, it refines the video search results list such that the top of the list better reveals intent. The third step further improves this refinement by visual reranking, exploiting intent-sensitive lightweight visual features extracted from thumbnails. Extensive evaluation of the approach includes a user study carried out on a crowdsourcing platform and a system-oriented evaluation. Evaluation results demonstrate that our approach leads to a substantial improvement of the information need satisfaction at users.**

***Index Terms*—Result optimization, user intent, video search.**



koi pond

Being A Fish In Japan - Koi and Koi Ponds

by SubmarineChannel • 3 years ago • 10,585 views

The coveted Koi carp is a symbol of male potency. And what our muscle-bound protagonist lacks...

Relaxing by the Koi Pond

by FISHGAL69 • 5 years ago • 138,884 views

Relaxing aquatics, fish feeding. Koi and butterfly koi.

how to build garden koi pond (step by step)

by phucle • 2 years ago • 55,720 views

how to build garden koi pond (step by step)

Fig. 1. Illustration of excerpts from a discontinuous video search results list. The list contains videos with several clearly focused dominant intents, but these are scattered throughout the list.

I. Introduction

The ultimate goal of a video search engine is to satisfy the user's information need in its widest scope. Numerous techniques developed for results lists optimization, e.g., [1]–[3], contribute to the current well-developed ability of video search engines to return search results that are on *topic*. On-topic videos do not, however, completely satisfy the user's information need unless they are also closely matched with the user's *intent*, i.e., the reason or purpose for which a user is searching for a video. Succinctly put, the topic component of the user's information need deals with *what* users are searching for, while the intent component deals with *why* they are searching.

This paper tackles the challenge of optimizing search results of video search engines so that they better capture the *why* aspect of the user's information need, while preserving the topical coverage requested by the query. We illustrate this challenge with the discontinuous results list depicted in Fig. 1. The user has submitted the query `koi pond`. With respect to topic, the precision of this list is perfect. However, for this query, the search engine returns videos that satisfy multiple categories of user intent: the first video best fits the goal of *acquiring information*, the second of *being entertained*, and the third of *learning a skill*. We refer to the intent class for which a given video appears to provide the best fit as the *dominant intent* of the video. Fig. 1 provides an illustration of how videos with different dominant intents occur scattered throughout the list and of the need to move the videos with the right intent higher up in the list.

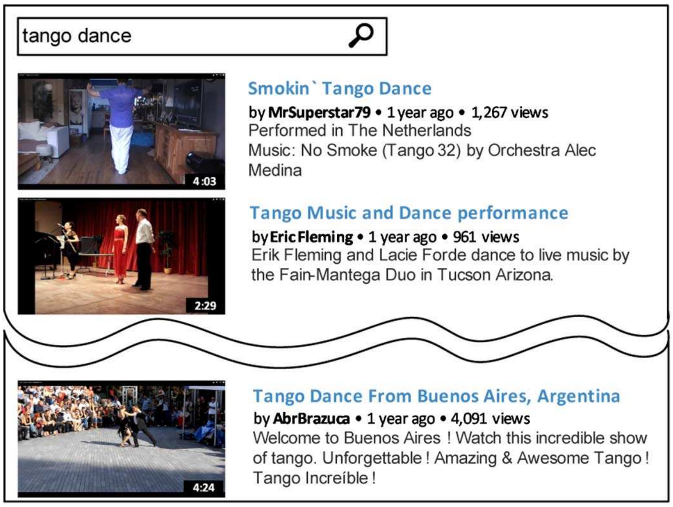
Analyzing the occurrence of dominant intent in video search results lists reveals that the intent distribution pattern varies strongly across different queries: Every results list follows one of two possible characteristic patterns where intent-aware optimization is needed. In the first case, videos satisfying different intents are scattered throughout the topically relevant results in the upper regions of the list (example in Fig. 1). In the second case, videos with a single dominant intent are prevailing in these regions. However, the videos that would be most helpful to the user because they are most clearly focused towards this intent may be buried relatively deep in the list. The latter case is illustrated in Fig. 2, which shows a discontinuous excerpt from the results list returned for the query `tango dance`. With this query, the user has decided to search for a video with the goal of *having a particular experience*, i.e., witnessing the Argentinian tradition of tango dancing. While the initial results list already covers videos satisfying the right intent, not all videos in the list match this intent to the same extent. Although all three videos

Manuscript received December 12, 2013; revised March 23, 2014; accepted March 24, 2014. Date of publication April 04, 2014; date of current version July 15, 2014. This work was supported in part by a Google Fellowship and in part by the Dutch national program COMMIT. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jiebo Luo.

The authors are with Delft University of Technology, Delft, The Netherlands (e-mail: c.kofler@tudelft.nl).

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/TMM.2014.2315777



tango dance

Smokin` Tango Dance
by MrSuperstar79 • 1 year ago • 1,267 views
Performed in The Netherlands
Music: No Smoke (Tango 32) by Orchestra Alec Medina
4:03

Tango Music and Dance performance
by Eric Fleming • 1 year ago • 961 views
Erik Fleming and Lacie Forde dance to live music by the Fain-Mantega Duo in Tucson Arizona.
2:29

Tango Dance From Buenos Aires, Argentina
by AbrBrazuca • 1 year ago • 4,091 views
Welcome to Buenos Aires ! Watch this incredible show of tango. Unforgettable ! Amazing & Awesome Tango ! Tango Increíble !
4:24

Fig. 2. Illustration of a discontinuous video search results list where videos that are best representative of the prevailing intent in the video search engine response are buried in the list.

show a tango being danced, only the third illustrated video, positioned rather low in the results list, portrays the traditional tango dance as practiced in Argentina and can best fulfill the user's original intent.

The challenge stated above is based on the assumption that users expend only a minimal amount of effort formulating their queries. Also, while they are relatively skilled in expressing the topical component of their information need in a query, they often fail to express their intent clearly, if they attempt to do so at all. It is therefore not surprising that topic rather than intent would be better reflected in the query string and that it is difficult to infer intent directly from the query. This reasoning motivates, a priori, the approach that we propose to tackle this challenge. In this approach we namely go beyond the query and exploit the structure of the results list itself. As illustrated by the two cases above and as will be demonstrated empirically later in this paper, videos are distributed over results lists in characteristic intent patterns. Our approach envisions refining the initial results list in view of these patterns. A novel approach in this respect is necessary because techniques that have been previously proposed to refine video search results lists topically, e.g., [1]–[3], cannot accommodate such patterns. Note that, since the topical component of an information need is largely independent of its intent component, we expect the intent patterns and the quality of our approach to be independent of topical relevance in initial lists.

II. Rationale and Contribution

Our work is inspired by previous work that established the importance of intent for users in video search and also showed that intent is reflected in both textual and content features of videos [4]. However, until now, the problem of intent-aware video search has not been tackled.

The approach proposed in this paper is based on the insight that results lists already contain many videos that are well-suited to support users in attaining specific goals, and are thus relevant to user's intent. As illustrated in Fig. 1 and Fig. 2, the challenge is not that these videos are missing from the results list, but rather that they are either scattered over or buried deeply and require difficult digging. In short, a query's results list contains a basic *intent response*, videos that potentially fit well with intent, but are not at the top of the list. A key characteristic of the approach is its ability to optimize video search results lists for two different categories of intent response patterns described above, that we will refer to in this paper as the *multi-intent* (Fig. 1) and *mono-intent* case (Fig. 2). While in a multi-intent case we aim at providing the optimal selection of videos in the top of the results list that reveal the most likely intents, in the mono-intent case we focus on pulling the videos that best match the prevailing intent to the top of the list.

In this work, we attack the challenge by analyzing the intent response reflected in the initial video search results list returned in response to a query. Even if it is not possible to infer the intent of the user who issued the query, we believe the user can still be supported by gathering a selection of videos having the highest potential to satisfy the user's intent from the topically-relevant part of an initial results list and by positioning them at the very top of the list. We adopt the typology of three basic intent classes relevant to video search as proposed by Hanjalic *et al.* [4]. This work provided basic evidence that videos satisfying similar intents share similar characteristics with respect to the vocabulary and language style used in their metadata and regularities detectable in their visual channel. We therefore exploit both textual and visual features extracted from the initial video search results list.

An important aspect of our approach is the way in which we represent the 'intent' of a video. Because a given video may provide a certain level of fit with a wide range of user's intents, we represent the intent of a video with a set of values that capture the extent to which the video fits each intent class in our underlying set of intent classes. However, when discussing the intent distribution patterns in video search results lists, we focus on the intent class that appears to provide the best fit with a video and considered it to be a video's 'dominant intent'.

Our approach consists of three steps. The first step analyzes the search engine response to a query from the topically-relevant part of an initial results list and classifies it as either mono-intent or multi-intent. The second step refines the search results lists in a way that it concentrates the videos having the highest potential for intent satisfaction at the very top of the results list. In the mono-intent case, it *reranks* the initial list in such a way that videos having the highest potential to satisfy the prevailing intent arrive at the very top of the optimized list. In the multi-intent case, where it is not possible to determine which of the many intents found in the initial results list is the one of the user, we focus on ensuring that diverse videos best representing different intents are present at the top of the list, i.e., we focus on optimally *diversifying* the initial list with respect to the returned intent response. The optimized rankings are produced by applying weighting parameters learned for both mono- and multi-intent lists. Using these weights, our algorithm is not only incorporating both optimization approaches into one common model, it is also not overly sensitive to potential intent response misclassifications from the first step and would still provide legitimate optimized results in these cases. The third step further improves the refined list with intent-aware visual reranking.

We evaluate our approach using both system-oriented evaluation and intent relevance judgments obtained through a crowdsourcing user study. Rather than using diversity metrics [5] designed with the goal of taking both ranking and diversity into account (and thus being unable to distinguish between mono- and multi-intent response optimization), we present our experimental results using standard evaluation metrics. Since, in this paper, we assume that upper regions of initial results lists produced by video search engines are topically relevant, we deliberately do not apply well-established concepts like query clarity [6] generally used to measure the consistency of results lists on a topic.

This paper makes three key contributions:

- An analysis of intent responses of a mainstream video search engine, i.e., YouTube, to a large number of queries corresponding to real-world user information needs. The analysis provides insight into typical patterns of intent distribution in video search results lists.
- An approach for intent-aware results list optimization that handles mono-intent cases and multi-intent cases in one integrated model and exploits the benefits of intent-aware visual reranking.
- A large user study, making use of a crowdsourcing platform, that demonstrates that intent-aware results list refinement and intent-aware visual reranking significantly improve the ability of video search engines to satisfy users' information needs.

The remainder of the paper is organized as follows. In Section III, we cover related work and discuss how it is exploited and extended by our proposed approach. Section IV reports the results of an analysis of intent distribution patterns in video search results. Our approach, which builds on these results, is presented in Section V. In Section VI, we present the evaluation of our approach and perform further analysis and discussion of our results in Section VII. We conclude the paper in Section VIII with an outlook on future work.

III. Related Work

In this section, we provide an overview of work most closely related to ours. Most generally, these are approaches that have been proposed for optimizing search result lists. We then go outside of the domain of video retrieval to look in general at intent-aware search, as it has been studied in other domains distinct from video search. Finally, we cover work that concerns evaluation techniques that make use of crowdsourcing.

A. Search Results List Optimization

Approaches that have been proposed in the literature to optimize search results lists typically either perform reranking (i.e., increasing the homogeneity of the top ranks of results lists in terms of relevance topics) or diversification (i.e., increasing the heterogeneity of the top ranks of results lists in terms of subtopics or individual topic aspects). In textual Web search, Wu and Crestani [7] merge multiple results lists in an unsupervised fashion following late fusion to obtain a reranked list by applying a linear combination of document scores. Approaches that have been applied to multimedia [8], make use of classification [1], clustering [9], [10], and graph-based techniques [2], [11]. These approaches are specific to multimedia because they exploit diverse modalities and sets of features. Agrawal *et al.* [12] approach results diversification by assigning higher relevance to documents sharing a high number of query classes with the query. In multimedia search, Liu *et al.* [13] mine relevant patterns from image search results of multiple search engines, exploiting the fact that the combination of different models can improve retrieval performance better than any of the individual models. Further, class-dependent retrieval models optimizing search results based on a particular (query-)class have been proposed. Yan *et al.* [14] classify queries into a set of target classes to deploy query-class-dependent retrieval functions fusing multiple modalities with class-optimized weights. Kennedy *et al.* [15] automatically discover query classes reflecting similar fusion strategies for multiple modalities and perform query-class-dependent retrieval. These query classes, however, focus on the topical dimension of a query and are unrelated to intent.

Our approach analyzes the intent pattern in a results list in order to infer its intent response category, and creates an optimal results list by combining several intermediate intent-specific results lists. These lists are fused using intent-class-dependent weights, which are optimized on the development set, corresponding to our intent response categories. In this regard, our approach bears a resemblance to general class-dependent retrieval models [14], [15] and resembles data fusion approaches, e.g., [7], [13]. The novelty of our approach is that it incorporates the specific characteristics of intent patterns in video search results lists (i.e., the existence of mono- and multi-intent cases) into one single model suited particularly for intent.

Our approach also shares similarity with visual-feature based results list optimization and in particular with approaches that make use of a limited number of potential reranking candidates from the initial ranking. Yan *et al.* [16] exploit pseudo relevance feedback and select relevant and irrelevant samples from initial lists, which serve as training data for reranking models. Liu *et al.* [17] compare video search results in initial lists in a pairwise fashion and apply the optimal set of preference pairs for visual reranking. Yang and Hanjalic [1] learn a query-independent visual reranking model for queries using query-dependent reranking features in a supervised fashion. The novelty of our use of visual features in comparison with these approaches lies in the fact that we select specific reranking candidates in an intent-informed fashion, i.e., the reranking candidates that our approach considers are all associated with the same dominant intent.

B. Intent-Aware Search

The majority of the work that has been carried out on intent-aware search concerns conventional text search. Here, approaches build on a well-known intent typology introduced in [18], [19]. For example, Jansen *et al.* [20] derive features from Web search transaction logs for each intent class and apply supervised learning for classification. Lee *et al.* [21] use features from users' click behavior in search sessions and anchor link distributions in a supervised learning approach. Cao *et al.* [22] use queries and their corresponding click-through data from search sessions as context information for queries to be classified. The work of Santos *et al.* [23], in the domain of textual Web search, has interesting parallels with our own. They propose a supervised intent-aware search result diversification approach by learning when to apply particular retrieval models for specific aspects of a query. The novelty of our approach, however, is that it concentrates intent at the top of the optimized results list for both mono-intent and multi-intent cases in video search, where [23] solely focus on the multi-intent case in the domain of textual Web search, apply the original intent typology from [18] and automatically classify a query into these classes before performing the optimization step. Azar *et al.* [24] define models for reranking search results with the objective to satisfying particular user types such that the most relevant search results are ranked at the top of the optimized results list. While this approach can be considered intent-aware, they do not take explicit intent information as originally defined in [18] into account.

When moving from one domain to another, it is necessary to develop new models of intent. The underlying difference between video-, image- and text-based search is well documented [4], [25], [26]. Kofler *et al.* [27] apply intent-specific results view adaptations for image search. Although this approach is intent-aware, it does not optimize lists such that documents with the highest potential for intent satisfaction receive higher relevance. Zha *et al.* [3] propose an approach supporting users to overcome their ‘intention gap’ by providing both keyword and image suggestion for an initial query. Cui *et al.* [28] rerank Web images based on a query image and the inferred ‘user intent’. Zhang *et al.* [29] exploit semantic descriptions and properties of visual concepts (e.g., ‘round’, ‘metallic’ etc.), link them in a hierarchic fashion to obtain better representations for images and collect user feedback in an online search session to produce refined rankings based on the user’s ‘intent’. However, these approaches define ‘intent’ as synonymous with ‘information need’ and, in contrast with our approach, they do not deal with the underlying goal of the search, but rather with the topical component of the query. Intent-aware optimization approaches have not yet been proposed for video search. Hence, our approach bears more relevance to papers that interpret intent as being one component *orthogonal* to other components such as topic [30].

Our approach builds on a user intent typology of three intent classes specifically derived from real-world user information needs in *video search* and a corresponding video intent classifier [4] to determine the extent to which each video in a results list fits each of the three intents. To the best of our knowledge, this is the only intent typology and classifiers explicitly developed for video search. Unlike [20], we deliberately do not infer the user’s intent directly from the query, but rather exploit the intent distribution in the initial ranking to classify a query into the two intent response categories introduced above. Our approach makes use of supervised learning to train a classifier, which can be applied to the intent response of the engine produced by unseen queries.

C. Crowdsourcing for Relevance Evaluation

In textual Web search, crowdsourcing has been exploited to label data and to obtain relevance judgments for relevance evaluation. Alonso *et al.* [31] and Eickhoff *et al.* [32] evaluated information retrieval systems using crowdsourced relevance judgments. Alonso and Mizzaro [33] compared judgments gathered from TREC assessors with corresponding crowdsourced relevance judgments and identified that crowdsourced annotations are similar to those collected from TREC assessors and that crowdsourcing workers detected errors in relevance judgments by experts. Blanco *et al.* [34] show that crowdsourced search system evaluations are repeatable over time and maintain stable and reliable results and confirm that results are comparable to expert judges. In our evaluation, we perform a large crowdsourcing user study to collect relevance judgments for baseline and optimized results lists with respect to intent.

IV. User Intent in Video Search Results Lists

Here we present the results of a qualitative analysis of the intent response patterns observable in a large number of real-world video search results lists. The objective of this analysis is threefold: it establishes the existence of the mono- and multi-intent response categories, shows that results lists produced by state-of-the-art video search engines can benefit from an approach to optimize them with respect to user intent and demonstrates the potential of visual thumbnails to reflect intent in results lists. This analysis and the investigations carried out in this paper build on a dataset comprising a total of 692 queries corresponding to real-world user information needs and ranked lists of video search results returned in response to these queries from YouTube. We emphasize, however, that the approach in this paper is developed without any particular bias towards a specific video search engine and that YouTube is chosen as we take it as representative of current online video search engine technology. Specifically for the analysis reported in this section, we create a development set and randomly select 100 queries from our dataset for this purpose. A detailed description of the dataset is provided in Section VI-A.

For our approach, we adopt an established user intent typology of three basic intents [4]: *Information*—users aim to obtain declarative knowledge, i.e., obtain information; *Experience*—users aim to obtain performative knowledge, i.e., acquire a skill, or have particular experiences of an actual person, place, entity or event; and *Affect*—users aim to change their mood or affective state, i.e., be entertained. It should be noted that our approach is not innately dependent on this typology and the chosen number of intents. Any other suitable intent typology and any number of intent classes could be easily adopted instead. We make use of a video intent classifier [4] which exploits that same intent classes share similar characteristics in the type and style of textual metadata associated with videos as well as their visual content. The classifier, trained using standard Support Vector Machines on the publicly available dataset [35] and exploiting multimodal information sources associated with the video, determines the extent to which a video fits each of the three intents and subsequently assigns the dominant intent to the video. In previous experiments [4] this classifier has been shown to significantly improve over baseline approaches. We therefore believe that deploying this classifier provides a sufficiently solid input into subsequent steps of our approach.

1) Video Search Engine Intent Response: For each query in our development set, we investigate the intent response covered in its initial results list, i.e., we apply the video intent classifier to each video in the initial results list and investigate the

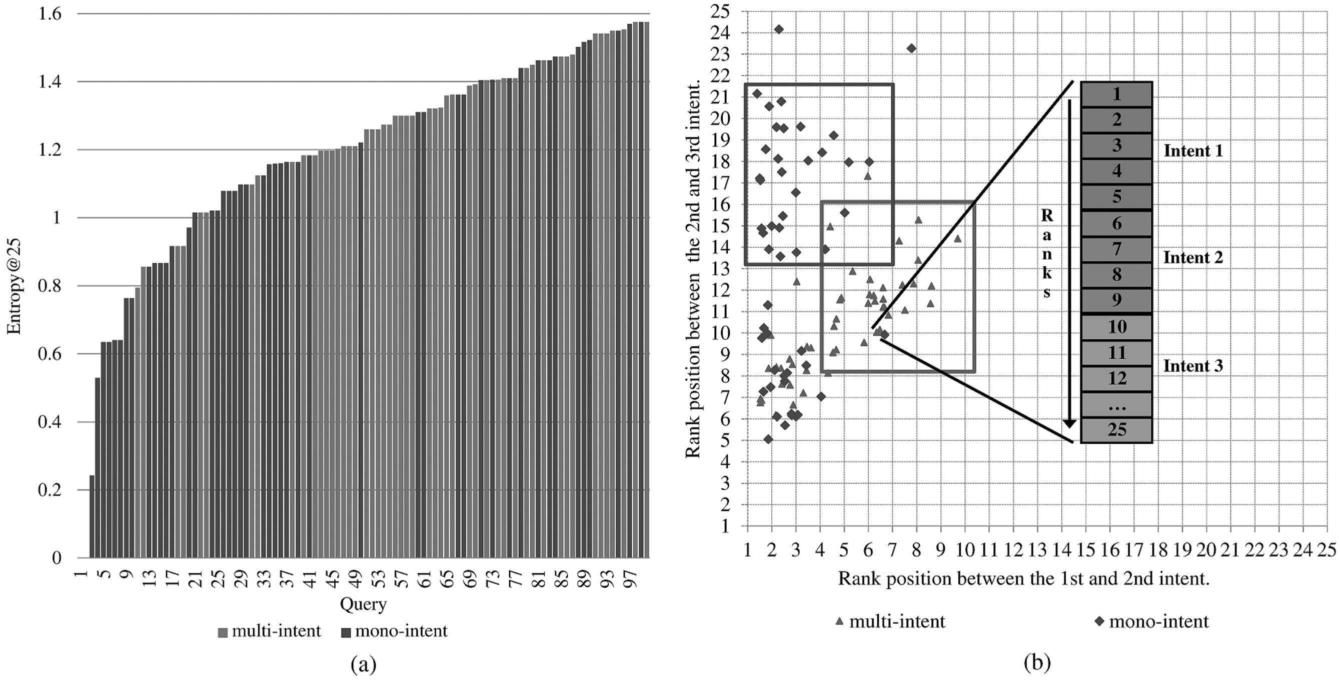


Fig. 3. (a) shows the entropy of intent classes among the top-25 video search results lists produced by the queries in our development set: mono-intent lists primarily show low entropy and multi-intent lists tend to have higher entropy. (b) shows the results lists as points determined by the ranks in the lists where the first video of a new dominant intent appears when following the list in a top-down fashion.

emerging intent class distribution. We are not interested in the low-ranked videos in the search results lists, since they typically show less relevance to the query and are therefore not topically optimized. Based on previous work on investigations of implicit relevance judgments of users [36] and informal observations of our development set, we work with top-N search results and choose $N = 25$ as the cut-off beyond which the topical relevance of videos drops off sharply, making them inappropriate for intent-aware optimization. We perform a crowdsourcing-based annotation process and ask users to assess the results list for each query from our dataset from the perspective of each of the two intent response categories, i.e., whether it satisfies one prevailing intent (i.e., mono-intent response) or multiple intents (i.e., multi-intent response). For each list, we then calculate the entropy among the emerged intent classes. Fig. 3(a) visualizes this investigation; the 100 queries of our development set are sorted by increasing intent entropy and are coded in two different shades of gray according to the ground truth label assigned to them in our crowdsourcing study. We clearly observe that the majority of mono-intent-response queries can be found on the left side of the plot, indicating low entropy, i.e., that the better part of search results of these queries share the same dominant intent. The majority of multi-intent-responses, on the other hand, show evidence of higher entropy, i.e., a balanced coverage of dominant intent classes covered in the search results. Please note that the difference between the two intent response patterns is not the absolute number of intent classes covered in the initial topically-focused results lists, but rather the degree to which each intent class is represented in the results lists.

We now look more closely at the individual initial results lists for the queries from our development set and investigate how the intents are actually distributed in the top-25 results. For each list, we start at the top of the list and record the rank position at which the first video with a yet unseen dominant intent appears. The results of this investigation are visualized by the scatter plot in Fig. 3(b). Note that the plot contains only those cases (96 queries) in the development set that manifest all three dominant intents and that multiple samples falling at a single point are slightly displaced for readability. An example in Fig. 3(b) illustrates how to interpret this plot: For a query plotted at the point [6/10], a possible intent distribution in the list could be that the first five ranked videos have the same dominant intent, the first video with a different intent appears at rank 6 and the first video with yet another unseen intent appears at rank 10. The distribution of the points in Fig. 3(b) reveals the variety of intents in the topically-focused top-25 region of the search results and shows that the patterns of intent distribution varies significantly across queries. We again recognize our two categories of intent responses: the mono-intent category (dark rectangle in the upper left part in the plot) covers queries for which initial lists contain videos that reveal one prevailing intent and the multi-intent category (light rectangle in the middle right part in the plot) covers queries for which initial lists contain videos that satisfy all intents relatively equally in the top-25 list. However, we observe that for a large number of queries the ranking with respect to intent is not optimal: in the mono-intent category the prevailing intent of the query may not be the first one to be found in the initial results list. In the multi-intent category, possible relevant intents other than the first one covered in the top of the results list may appear too late and in suboptimal balance when following the list in a top-down fashion.

2) Intent Response and Visual Thumbnails: Over the entire indexed video collection, a single intent class can be expected to be associated with a wide variation in visual appearance of the video. However, the visual optimization step in our approach does not operate on the entire video collection, but rather on a results list that is already constrained in the sense that each video is topically relevant to the query. Under such circumstances, intent can be expected to be more visually stable. In other words, videos with the same dominant intents will be more visually similar to each other than they are to videos with different dominant intents in the same list. We investigate whether inferring such similarities from the thumbnails representing a video can benefit visual reranking and provide further improvement. For all thumbnails in the initial lists produced by the queries in our development set, we extract visual features (see Section VI-A for details) and represent each thumbnail in the form of bag of visual words [37]. Inspired by the concept of query clarity [6] and how it was applied in multimedia search [38], we build a language model for each intent class in each results list, i.e., each model comprises the videos satisfying the same dominant intent. We calculate the Kullback-Leibler (KL) divergence between each intent class language model to determine how coherent these models are and perform a Wilcoxon signed-rank test (condition: $p < 0.005$) on the KL divergences. The results indicate a significant difference between classes *Information* and *Affect* ($p = 0.004621$) and *Experience* and *Affect* ($p = 0.004219$); presumably due to a higher level of similarity between *Information* and *Experience*, no significant difference was observable here ($p = 0.00627$). This fact leads us to expect that an intent-informed selection of thumbnails provides a weak indicator, which, however, we assume to be strong enough to be beneficial for our approach.

V. Approach

Here, we describe our intent-aware video search result optimization approach. It consists of three steps. In the first step, we automatically analyze the search engine response and classify it as either mono- or multi-intent (Section V-A). In the second step, we optimize the initial ranking in a way that it concentrates the videos having the highest potential for intent satisfaction at the top of the results list depending on the predicted intent response (Section V-B). In the third step, we further improve the refined list by performing visual reranking, which selects potential visual reranking candidates from the initial ranking in an intent-informed fashion (Section V-C). Fig. 4 gives a scheme of our approach, which includes an offline processing pipeline to extract intent response features and results lists optimization features for training and an online processing pipeline, where the trained models are applied to unseen queries.

A. Intent Response Classification

Recall that we apply the video intent classifier from [4] to determine the confidence with which each video in an initial list satisfies each intent. Then, each video is represented by a vector indicating how well it satisfies each intent. The length of the vectors corresponds to the number C of intent classes; here, we work with $C = 3$. Each row in Table I exemplarily illustrates these vectors for videos ranked 1 through N. To determine to what extent each intent class is reflected in the engine's overall response for an initial list, we merge the intent confidences from all top-N videos into an *intent response vector* $\mathbf{Q}$; the merging process is performed by building the median value of each intent-based list (cf. intent-specific columns in Table I).

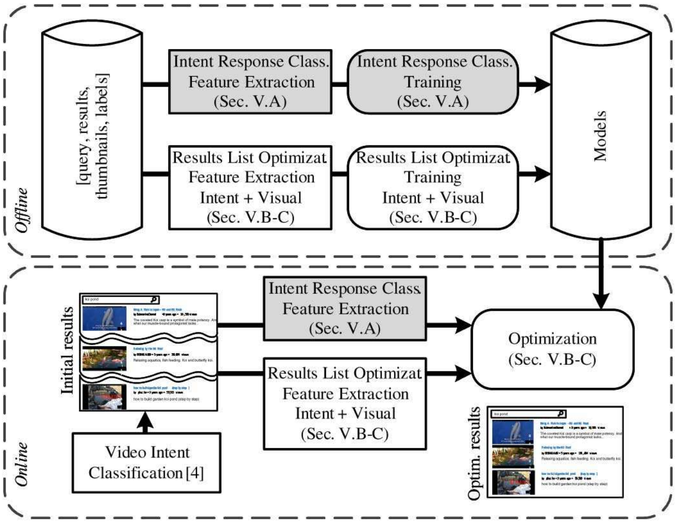

[query, results, thumbnails, labels]
Intent Response Class. Feature Extraction (Sec. V.A)
Intent Response Class. Training (Sec. V.A)
Models
Results List Optimizat. Feature Extraction Intent + Visual (Sec. V.B-C)
Results List Optimizat. Training Intent + Visual (Sec. V.B-C)
Offline
Initial results
Intent Response Class. Feature Extraction (Sec. V.A)
Optimization (Sec. V.B-C)
Results List Optimizat. Feature Extraction Intent + Visual (Sec. V.B-C)
Video Intent Classification [4]
Optim. results
Online

Fig. 4. Intent-aware video search result optimization approach overview.

TABLE I
Illustration of How to Build Intent Response Vector $\mathbf{Q}$

Initial Rank	**Video intent classifier confidence scores**		
	Information	*Experience*	*Affect*
1	0.20	0.60	0.20
2	0.33	0.33	0.33
⋮	⋮	⋮	⋮
N	0.60	0.20	0.20
$\mathbf{Q}$	0.33	0.33	0.20

The intent response vector is designed to capture information used to automatically distinguish between mono- and multi-intent responses. We expect intent response vectors for queries producing mono-intent responses to explicitly point to one distinct dominant intent satisfied in the search results (e.g., 0.50, 0.20, 0.20), and multi-intent responses to not clearly reveal any particular intent (e.g., 0.33, 0.33, 0.20). We model this distribution with the variance among the components of the intent response vector, $\sigma^2(\mathbf{Q}) = \frac{1}{C-1}\sum_{i=1}^{C}(Q_i - \mu)^2$, with Q_i and μ being vector components and their mean value, and we expect higher variances among components of mono-intent vectors and lower variances for multi-intent vectors.

To automatically distinguish between the two intent response categories in the results lists of unseen queries, our approach makes use of supervised learning, where the intent response vector $\mathbf{Q}$ and the variance $\sigma^2(\mathbf{Q})$ among its components serve as features to train the classifier in an offline step. We adopt a rule-based classifier, known to perform well on short feature vectors. Note that the features we apply are query-independent, making the approach applicable to unseen queries: For a new query, our approach automatically builds $\mathbf{Q}$ and $\sigma^2(\mathbf{Q})$ and feeds these features into the trained classifier to predict the query's intent response class. Because all time-sensitive operations are performed during training, the computational costs during query time are negligible.

B. Intent-Aware Results Lists Optimization

Once we classified a query's intent response into one of the two classes, we proceed with refining the initial list accordingly.

In the mono-intent case, our approach focuses on *reranking* the initial list in such a way that videos having the highest potential to satisfy the prevailing intent are on the very top of the optimized list. 'Highest potential' is defined as videos for which the video intent classifier assigns a higher probability to the prevailing intent class of the video than to other videos in the initial list. In the multi-intent case, our approach focuses on ensuring that diverse videos best representing different intents are present at the top of the list, i.e., it focuses on optimally *diversifying* the initial list with respect to the intent response of the initial list. We refer to these approaches as RR and DV.

We adopt an optimization model based on [1] and incorporate both optimization approaches into one common model. Similar to [7], [13], [15], our optimization approach follows a late fusion principle where we merge the initial ranked video search results list $\mathbf{R}$, defined as a vector of descending ranks, and several alternative results lists generated according to different criteria together in a final, optimized ranked list of video search results. To generate the alternative lists, we rerank the initial list $\mathbf{R}$ in view of each considered intent class according to the descending confidence scores generated by the corresponding video intent classifier (reordered values in the intent-specific columns in Table I). If two videos match in score, the original ordering is maintained. This leads to *intent-basedrerankedresults lists* $\mathbf{I}_i$ $(1 \leq i \leq C)$.

The optimized ranking is produced by reordering the initial list based on the score vector that is computed by fusing the score vectors of the initial ranking $\mathbf{R}$ and of the intent-based rankings $\mathbf{I}_i$ using linear combination where we steer whether to rerank (for mono-intent cases) or diversify (for multi-intent cases) with response category-dependent weights parameterizing our model. Using our model, each video v from an initial results list gets an updated relevance score S_v assigned. Please note that we refer to the rank of a video v in the initial ranking $\mathbf{R}$ with R_v (and likewise for other results lists). We formulate the model as

$$S_v = \lambda \cdot s(R_v) + (1 - \lambda) \sum_{i=1}^{C} \tau_i \cdot s(I_{i,v}) \tag{1}$$

where λ and τ_i are parameters steering the importance of the initial ranking ($\lambda = [0, 1]$) and the rankings produced by the intent-based lists ($\sum_i^C \tau_i = 1$). Similarly to [7], we use function s to assign a score to each video according to its ranking in a particular list:

$$s(r) = \frac{1}{N}(N - r + 1). \tag{2}$$

The most straightforward way to refine the initial results list and set parameters λ and τ_i would be to directly apply the individual components of the intent response vector $\mathbf{Q}$ of a query as weighting parameters τ_i without utilizing supervised learning. We conjecture, however, that these parameters would only be suboptimal and not exploit the full potential of our approach, since they rely on the intent distribution of one particular query and are not optimized with respect to our intent response categories. In our experiments we will demonstrate that optimizing parameter settings through supervision outperforms this naïve approach.

Thus, we follow a procedure building on supervised learning to optimize the weighting parameters λ and τ_i for our two intent response categories. Similarly to [15], for all queries in our development set, we perform a full grid search of all possible combinations of λ and τ_i fulfilling their boundary conditions (step size: 0.1). After obtaining an optimized list for a query with a particular parameter combination, we build an *optimized intent response vector* $\mathbf{O}$ for the top N_O videos in the optimized results list ($C \leq N_O \leq N$), which we use as input to minimize the loss function

$$l(g, \mathbf{O}) = 1 - \left[\left(\frac{1}{\sigma^2_{max_C}} \right)^{(1-g)} \left| g - \sigma^2(\mathbf{O}) \right| \right] \tag{3}$$

where g presents the intent response ground truth label in terms of mono-intent ($g = 0$) or multi-intent ($g = 1$) for a query. $\sigma^2_{max_C}$ is the maximum variance for a vector with C components (e.g., the maximum possible variance of a vector with $C = 3$ components is $\sigma^2_{max_C} = 0.3\dot{3}$) used for normalization in case of intent responses corresponding to mono-intent. The idea of the loss function is straightforward and similar to our intent response classification approach: it exploits expected high variances among components of the optimized intent response vector for mono-intent and expected low variances for multi-intent responses. l reflects this expectation as it converges to 0 if the intent variance among the top N_O optimized results reflects the expected intent distribution, i.e., videos best satisfying the prevailing intent for the mono-intent case and videos satisfying multiple intents in the multi-intent case did move to the top N_O results. In other words, the procedure optimizes the parameters λ and τ_i in such a way that the intent response ground truth label is reflected by the intent response of the top N_O results of the optimized ranking. The parameters are learned and optimized during training and can then be efficiently applied to rankings produced by unseen queries in real-time with insignificant computational costs.

C. Intent-Aware Visual Reranking

Our qualitative analysis from Section IV-2 lets us expect that a further improvement of the intent-aware results lists is possible by exploiting *visual similarity* among thumbnails. We extend our RR and DV approaches with an intent-informed selection process for thumbnails benefiting visual reranking and refer to this extension with RR + VRR and DV + VRR. Inspired by [1], [2], [16], for each thumbnail (i.e., the *target thumbnail*) in an initial list, we build features reflecting its neighborhood structure. We calculate the thumbnail's visual similarity to a set of thumbnails from the same results list. Compared to the approaches using *all other* [1] or *pseudo-selected* [16] thumbnails from the same list as visual neighbors, our approach is more selective as it solely considers thumbnails belonging to videos which share the same dominant intent as the target thumbnail: Given a video v from the initial results list and having the dominant intent class i, we obtain its visually similar search results by computing its visual neighbors $\mathbf{V}$ from the set of videos in the results list having the same intent class as video v. Thumbnails in these sets are additionally solely considered as neighbors if their similarity to v is larger than a threshold ϵ. This more selective approach will leverage topic-specific visual similarities between videos with a given dominant intent, which is expected to sharpen the intent focus at the top of the list.

In order to assign a reranking score R_{visual} to each video, we utilize its visual context information $\mathbf{V}$ by performing soft neighborhood voting [1]. This voting principle—here tailored towards user intent—exploits for each video v its visual neighbors $\mathbf{V}$ and their ranks in the respective intent-based results lists $\mathbf{I}_i$, assigning a larger importance to visually similar thumbnails of videos higher ranked in the list $\mathbf{I}_i$:

$$R_{visual} = rsv(\mathbf{V}) = \sum_{t \in \mathbf{V}} \frac{1}{\log(I_{i,t} + 1)}. \tag{4}$$

The final visually reranked list is produced by reranking the videos by descending ranking scores R_{visual}. Since the generic optimization model presented in Eq. (1) builds on late fusion, we can incorporate the now produced visually reranked list by adding it to the right part of Eq. (1) with its own weighting parameter τ_{C+1} and by repeating the weighting parameter optimization such that condition $\sum_i^{C+1} \tau_i = 1$ is fulfilled. Since features are extracted during indexing in a computationally inexpensive manner and solely thumbnails sharing the same dominant intent category are compared during query time, the critical computational costs are low, making our approach applicable for large-scale video search.

VI. Experiments

A. Experimental Setup

1) Dataset: To properly evaluate how well our approach optimizes results lists from an intent-aware perspective it is critical to take the explicit information need behind a query into account. Our dataset comprises a total of 692 video search information need requests gathered from Yahoo! Answers and their corresponding queries [39]. The information need requests were filtered and the queries were derived from the information need statements using a crowdsourcing-based procedure involving three workers per statement and carried out on Amazon Mechanical Turk (AMT): We asked workers to create a short keyword query that could help to find the content the requester is looking for and merged the three queries to a final query per information need statement (on average, term overlap between worker-suggested queries was approximately 63%) [39]. For each query, we then collect initial results lists from YouTube. For our intent response classification as well as intent-aware results lists optimization, we restrict the lists to their top $N = 25$ relevant results returned by YouTube, since these results are considered being most topically focused [36] and our qualitative analysis showed that a good coverage of mono-intent and multi-intent response categories is already achieved on top-25 lists. Restricting our initial lists to the topically most relevant videos avoids the induction of videos that are not topically relevant into the list during intent-aware refinement. For each of the videos in the search results lists, we collect metadata including tags and YouTube category labels as well as the visual thumbnail that was selected to represent the video by the user who originally uploaded the video.

2) Visual Feature Extraction: To represent thumbnails for our visual reranking approach, we extract the low-level Fuzzy Color and Texture Histogram (FCTH) feature [40] from each thumbnail. This feature lets us not only experiment with an effective image representation, the computational costs are also relatively low, making it appropriate for use in online search sessions on large-scale data collections. The choice for lightweight features reveals our goal for this paper, namely to show, for the first time, the potential of deploying the visual channel of videos to sharpen the intent focus in results lists. We believe it is realistic to assume that better visual representation of the thumbnails, involving for example semantic visual concepts [41], should lead to more benefit of the visual channel to intent-aware video search results reranking. We leave, however, elaborate studies towards maximizing this potential for future work. The thresholding parameter ϵ controlling the cut-off for the final set of visual neighbors for soft voting for intent-aware visual reranking was empirically set to 0.75 [1].

3) Mono- and Multi-Intent Response Ground Truth: We use the $C = 3$ basic intent classes and the video intent classifier [4] to determine the extent to which each video in a results list fits each intent. To generate ground truth for our mono-intent (hereafter referred to as `mo`) and multi-intent (`mu`) intent response categories, we perform a crowdsourcing-based annotation process on AMT. We asked users to judge results lists produced by each query with respect to these classes, i.e., whether its results satisfy one prevailing intent or multiple intents [4]. Because the initial results lists are too long to consider all videos, we select three videos from the top-25 list as a representative sample [4]. 313 queries were annotated as `mo` and 379 as `mu`. For experiments, we remove the randomly selected development set (100 queries) used in our qualitative analysis, resulting in a test set comprising 592 queries (263 `mo`, 329 `mu`) and 13,024 videos for experimentation. This dataset will be made publicly available with this paper.

4) User Intent Relevance Judgments: User intent relevance judgments enable us to evaluate and analyze how well the initial and optimized results lists satisfy users' intents, and, ultimately, their information need. Note that our queries do not have any topical categories assigned, as they would not be useful for our evaluation. We obtain the intent relevance judgments by carrying out another AMT crowdsourcing process that collects user feedback about intent. For each query and information need in our dataset, AMT workers are presented with the information need statement, the corresponding query and two lists of top-3 search results, one from the baseline and one from the refined list. We focus on evaluating the quality of the top $N_O = 3$ search results of the optimized list. Such a choice is consistent with the results of studies on short-term human memory, which set its size to be 7 ± 2 [42]. By presenting workers with six items, we reduce the risk that their judgments are impacted by the difficulty of comparing a too-large set of items. We ask crowdsourcing workers to judge each list with respect to whether (i) one specific video best satisfies the intent reflected in the information need, (ii) this video better satisfies the intent than videos in the other list, and (iii) there is a second video in the list satisfying the user's intent. This information gives us the ground truth intent relevance. Specifically, we know in which list the video that best satisfies the intent of the information need occurs and we know how users perceive the relevance of other videos in the list related to that of the best video. Each query was annotated by three workers and we applied inter-annotator agreement using majority voting to fuse annotations of each query. We applied standard quality control practices. To ensure that workers judged videos and lists with respect to intent, we asked questions to help putting the workers in the position of the user who originally had the information need in mind. In total, 129 workers participated in the user study and inter-annotator agreement surpassed the generally accepted level of 0.7 with an average kappa of 0.739.

5) Evaluation and Baselines: We evaluate our approach by performing extensive experiments to answer the following research questions:

RQ1 Can the intent response of the video search engine be exploited to automatically distinguish mono-intent and multi-intent search engine responses?

RQ2 Can the proposed optimization approach lead to improved satisfaction of user's information need?

RQ3 Can pure intent-aware results-list optimization ($\mathtt{RR}/\mathtt{DV}$) be further improved through visual reranking ($\mathtt{RR}/\mathtt{DV} + \mathtt{VRR}$)?

To answer **RQ2** and **RQ3**, we perform both a system-oriented evaluation (Section VI-C) and analyze the intent-based relevance judgments obtained by our large crowdsourcing user study (Section VI-D). Algorithms building on supervised learning (used both for intent response classification and optimization parameter learning) are evaluated using 10-fold cross-validation. For system-oriented performance evaluation (both for classification and results lists optimization), we report our results in terms of Weighted F-measure (WFM), defined as F-measure (FM) weighted by the intent response category size (the number of queries in the respective intent response category in our dataset). To measure how well the optimized results satisfy users' intents, and, ultimately, their information need, we report our results in terms of Mean Reciprocal Rank (MRR) using the user intent relevance judgments obtained by our crowdsourcing user study. Improvement is expressed on a relative scale. Statistical significance tests were performed using the Wilcoxon signed-rank test ($p < 0.05$).

It is realistic to assume that more sophisticated solutions could be deployed for combining different candidate results lists as alternatives to the solution presented in Section V-B. However, as indicated by **RQ2**, the goal of this paper is not to find the best results list-aggregation strategy, but to show, for the first time, that the intent-awareness of video search results can be improved and that taking intent-based reranked results lists into account is a promising way to achieve this goal. We leave elaborate studies on maximizing list-combining effectiveness in this problem context and the application of other optimization approaches for future work. For this reason, we evaluate our optimization steps against the baseline performance represented by the initial results lists produced by YouTube as well as the unsupervised way to refine the initial list by directly applying the individual components of the intent response vector as weighting parameters.

6) Approach Pipeline Evaluation: Although ultimately the output of the intent response classification would feed directly into results list optimization, we chose to evaluate each step separately. We chose this setup, since the objective of this paper is to evaluate, for the first time, the potential of intent-awareness in video search and conjecture that separate evaluation of the two steps allows us to gain particular insight into the strengths and weaknesses of each. For the results optimization step, this insight is particularly valuable, since evaluation requires a large crowdsourcing study and for this reason a rather substantial investment of research resources. We calculated that at a maximum only 6% of the information needs could be impacted by our decision to evaluate the pipeline steps separately, and leave detailed investigation of this fraction for future work.

TABLE II
RESULTS OF OUR INTENT RESPONSE CLASSIFICATION APPROACH

Method	FM	FM	WFM
	mo	mu	
Dominant class baseline	0.000	0.707	0.387
YouTube category distr. baseline	0.447	0.631	0.548
Rule-based classifier	0.525	0.659	**0.598**

B. Intent Response Classification

We report on our experiments evaluating how well our intent response classification approach can automatically distinguish between mono-intent and multi-intent engine response, i.e., how well it classifies unseen queries in the response categories mo and mu. We compare our approach to the dominant class baseline, which reflects a scenario in which all queries are automatically classified in the larger of the two response categories (i.e., mu). We also implement another, more sophisticated baseline approach, where we exploit the distribution of YouTube categories (e.g., *Science & Technology*, *Sports* etc.) assigned to each video in initial results lists. We chose this baseline because the YouTube categories reflect the topic and style of videos, presenting an additional typology of categories not directly related to intent. Table II presents a performance overview in terms of WFM.

First, we observe that the YouTube category distribution baseline (0.548) statistically significantly outperforms the dominant class baseline (0.387). Second, our intent response classification procedure exploiting the intent response signal from initial lists results in a 9.1% improvement (0.598). This result was expected, and serves to demonstrate that our approach is indeed sensitive to intent classes and not to other general categories into which YouTube videos could be categorized. These results allow us to give a positive answer to research question **RQ1**: intent response of the video search engine can be exploited to automatically distinguish between the intent response categories mo and mu.

C. General Results Optimization Performance

We report on our experiments evaluating how well the optimized results lists fit the expected overall patterns of their respective response category. These results report on the performance achieved by the parameters trained on the loss function defined in Eq. (3) and show how well videos were reranked for mono-intent- and diversified for multi-intent responses. How well our optimized results truly benefit the user's intent satisfaction is reported in the next subsection. Table III presents a performance overview in terms of WFM. We compare our approaches to a set of baselines. The first baseline is presented by the top-3 results of the initial ranking produced by YouTube. The second baseline is the discussed

TABLE III
RESULTS OF OUR APPROACH EVALUATING HOW WELL THE OPTIMIZED LISTS FIT THEIR EXPECTED RESPONSE CATEGORY

Method	FM	FM	WFM
	mo	mu	
YouTube baseline	0.528	0.549	0.540
RR/DV unsupervised baseline	0.573	0.555	0.563
RR/DV	0.658	0.609	**0.631**
RR/DV + non-intent-aware VRR [1] baseline	0.655	0.609	0.629
RR/DV+VRR	0.680	0.638	**0.657**

straightforward unsupervised way to refine the initial list by directly applying the individual components of the intent response vector as weighting parameters τ_i. We compare our combined approach RR/DV + VRR additionally against another baseline, which, similar to [1], selects *all* visual thumbnails from each results list as visual neighbors for visual reranking. Note this condition does not select them in an intent-informed manner, i.e., it does not apply Eq. (4).

First, we observe that the unsupervised baseline approach (0.563) outperforms the YouTube baseline (0.540), however, only with a moderate improvement, indicating that a completely unsupervised optimization approach is restricted and that supervised parameter optimization is necessary. Second, all proposed supervised parameter optimization approaches statistically significantly outperform both baseline approaches. Applying pure intent-aware optimization RR/DV (0.631) results in an improvement of 16.9% over the YouTube baseline. This improvement shows the effectiveness of our pure intent-aware approach. Not only that our supervised approach overall produces better optimization results for mono- and multi-intent cases, applying the weighting parameters obtained through supervised learning also makes it less sensitive to possible intent response misclassifications compared to the baseline approach. Third, combining RR/DV with the non-intent-aware visual reranking baseline approach taking *all* visual neighbors into account [1] (0.629) decreases the system's performance. This outcome was expected, since the visual neighbors are not selected in an intent-informed fashion, introducing noisy results in the top of the visually reranked lists. Applying our combined, fully intent-aware RR/DV + VRR approach (0.657) results in a 21.7% improvement over YouTube. The 4.1% improvement over pure intent-aware refinement (0.631) reflects our expectations of moderate improvement using visual reranking. Importantly, it strongly supports the idea behind our visual reranking approach that within a topically-focused list intent is reflected by visual similarities, which can be exploited for selecting visual reranking candidates.

Table IV presents our optimized weighting parameters λ (initial list), τ_1 (*Information* list), τ_2 (*Experience* list), τ_3 (*Affect* list) and τ_4 (visually reranked list) trained on our development set. First, the τ_i parameters for mo are biased towards reranking with respect to the most probable intent list ($\tau_3 = 1$) and mu yields a more balanced weighting, diversifying the initial results. Second, visual reranking has slightly more influence on class mo ($\tau_4 = 0.4$) than for mu ($\tau_4 = 0.3$), implying that for mono-intent responses visual reranking has a larger effect to refine the top results.

TABLE IV
OPTIMIZED WEIGHTING PARAMETERS

Method	Class	λ	τ_1	τ_2	τ_3	τ_4
RR/DV	mo	0.6	0.0	0.0	1.0	—
	mu	0.6	0.1	0.7	0.2	—
RR/DV+VRR	mo	0.5	0.0	0.0	0.6	0.4
	mu	0.5	0.0	0.5	0.2	0.3

TABLE V
RESULTS OF OUR APPROACH EVALUATING HOW WELL THE OPTIMIZED LISTS SATISFY THE USER'S INTENT

Method	MRR	MRR	MRR
	mo	mu	
YouTube baseline	0.494	0.333	0.461
RR/DV	0.614	0.621	**0.617**
RR/DV+VRR	0.662	0.755	**0.713**

These results allow us to answer **RQ2** positively: initial results lists can be optimized to fit the expected patterns of their respective response category. Since weighting parameter λ for the initial ranking is relatively large, the topical focus of the query is not decreased. We can also answer **RQ3** positively: RR/DV can be outperformed by extending it with visual reranking, and, most importantly, solely in an intent-informed manner.

D. Intent Satisfaction Performance

We report on our experiments evaluating how well the optimized results lists satisfy the users' intents, and, ultimately, their information need. Using the judgments obtained through our crowdsourcing user study, we compare the top-3 results of the initial ranking from YouTube with RR/DV and in a second experiment the ranking produced by RR/DV with RR/DV + VRR. Table V presents an overview of our performance evaluation in terms of MRR. To calculate the reciprocal rank score per query for a results list, we use the ranking position of the video which the crowdsourcing worker selected as best satisfying the user's intent.

First, both optimization approaches statistically significantly outperform the baseline approach. In the setup comparing the initial YouTube baseline (0.461) with optimized lists generated by our pure intent-based approach RR/DV (0.617), we achieve an improvement of 33.8%. In the second setup, which compares RR/DV with its extension of intent-aware visual reranking RR/DV + VRR (0.713), we achieve an improvement of 15.6%. Both performance improvements show that the effectiveness of our approaches is positively influencing the user's intent satisfaction. Next to a large performance improvement using RR/DV, the improvement of RR/DV + VRR illustrates that only slight modifications in the rank within the top-3 search results contributed by intent-aware visual reranking are already appreciated by users with respect to their intent satisfaction. Second, RR/DV achieves a significant improvement for both classes over the baseline, however, the performance improvement for the multi-intent response category is larger (0.614 vs. $0.494 = 24.3\%$ for mo and 0.621 vs. $0.333 = 86.5\%$ for mu). This difference is expected, since our approach is motivated by the fact that users would receive a better intent satisfaction for queries which are expected to reflect multiple intents. However, this result also demonstrates that users recognize videos having the highest potential for intent satisfaction in the top of the

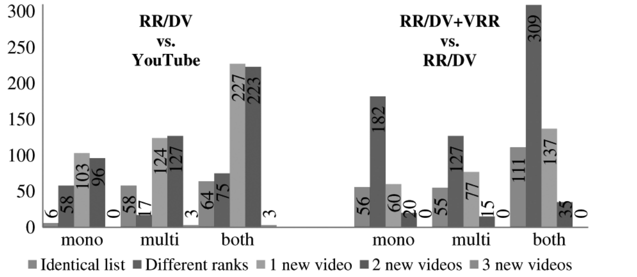


Fig. 5. Summary of how the top-3 results of optimized lists change compared to the baseline ranking.

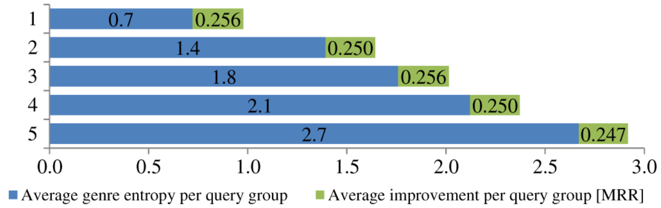


Fig. 6. Performance improvement of our RR/DV + VRR approach over the baseline for queries having different topic distributions in their search results.

list, regardless of the engine's initial intent response. Third, applying RR/DV + VRR also results in a larger improvement gap for the multi-intent case (0.755 vs. $0.621 = 21.6\%$) compared to the mono-intent class (0.662 vs. $0.614 = 7.8\%$).

These results allow us to answer our research question **RQ2** positively also from the perspective of users: our intent-aware optimization approach RR/DV clearly better satisfies the user's intent than the initial ranking. We can again give a positive answer to **RQ3**: the refinement introduced by our combined approach RR/DV + VRR is recognized by users.

VII. Analysis and Discussion

A. *Intent-Aware Results Lists Optimization*

To attain a better understanding of our proposed approaches and their strengths and weaknesses, we perform a thorough analysis of the obtained results. First, we analyze the lists produced by RR/DV and RR/DV + VRR and count the number of queries for which the optimized lists, compared to the lists of the baseline approach, are identical, contain the same 3 videos with a different ranking, or contain 1, 2, or 3 new videos. The analysis results are visualized in Fig. 5.

Referring to the left side of Fig. 5, RR/DV changes the initial list by introducing either one ('1 new video') or two ('2 new videos') new videos in the list. This result shows that for both the mono- and multi-intent case new videos are introduced in the top-3 ranking. However, these changes have slightly more effect on queries following a multi-intent engine response pattern, supporting our experimental results: Since for multi-intent cases the query does not have a single prevailing intent, more new videos moved upward from the initial list in the top-3 search results. RR/DV produces only 3 lists which are totally different compared to the baseline ranking ('3 new videos'); these changes only occur for the multi-intent pattern case. These observations also confirm that our approach does not reduce the topical focus of the query reflected in the initial ranking.

Referring to the right side of Fig. 5, RR/DV + VRR contributes most in optimizing the already refined top-3 ranking, where the optimized list contains the same videos as the already refined list produced by RR/DV, however, with a different ranking ('Different ranks'). We further observe that if one new video is introduced into top-3 ranks, it mostly affects multi-intent responses ('1 new video'), that many already refined lists do not change by further improving it using visual reranking ('Identical list'), and that no list was further refined such that three totally different videos were introduced ('3 new videos'). These observations are consistent with our expectations of the approach: namely, that visual reranking helps in particular cases, that it is a rather conservative reranking approach, and that it makes most of its contribution in terms of the refinement of the top ranks rather than introducing new videos at the top of the list.

Next, we investigate the influence of topical relevance in initial lists on the final performance of our approach. Our standpoint is that, although the search engine output is topically optimized, topical optimization is not enough to fully satisfy user needs and further improvement is needed. We expect that, ultimately, our approach will provide results that better satisfy users' information needs for initial lists containing more topically-relevant results. However, since information needs have both topical- and intent-components that are largely orthogonal to each other, we expect the quality of our approach to be independent of topical relevance in initial lists.

In order to investigate the interplay between topical relevance and intent relevance, we turn to the genre labels of the videos in results lists. Although genre is not directly related to topic, the relationship is tight enough to give us an interesting window on the behavior of our approach. If we assume that the query is topically related to a single topic, then we would expect results lists containing highly relevant results to also contain videos that are related to one or a very small number of genres. On the basis of this assumption, we hypothesize that results lists containing many different genres of videos more probably contain many videos that are not relevant to the query. We measure the variety of genres by calculating the genre entropy among each query's top $N = 25$ search results. We then group queries by their entropy scores in five equally-sized groups and, for each group, calculate the improvement in terms of MRR obtained by our approach compared to the baseline (cf. Fig. 6). This analysis provides a demonstration that our approach engages with the intent dimension of results relevance, and provides evidence of orthogonality with the topical dimension. This analysis supports our claim that the ability of our approach to improve the intent relevance of a topic list is effectively independent of the level of topical relevance of the initial list to the query.

B. *Individual Query Performance Evaluation*

We analyze individual queries and how their optimized lists differ from their initial lists. This analysis particularly focuses on demonstrating how our reranking and diversification approaches were applied to mono-intent and multi-intent engine responses and how their combination with intent-informed visual reranking performs. Fig. 7 gives examples for optimized lists, which were selected by calculating the performance improvement achieved by our approaches over the best-performing baseline approach.

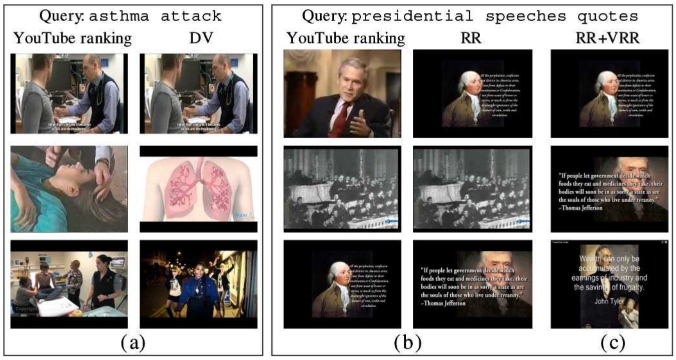


Fig. 7. Top-3 results for example queries. (a) and (b) present the YouTube baseline ranking and the pure intent-optimized ranking for a multi- and a mono-intent response query, respectively. (c) presents the optimized ranking of (b) further refined by intent-informed visual reranking.

Query `asthma attack` follows a typical multi-intent engine response pattern: Although the initial top-3 results contain videos covering solely one intent—to experience how to react in case an asthma attack happens—videos satisfying other intents are scattered over the top-25 initial results. Since this query is a multi-intent query, our approach diversified the list to ensure that diverse videos representing different intents are at the top of the list (`DV`): Video 1 lets users experience how to react in case an asthma attack happens, video 2 serves to inform, and video 3 is an entertaining music video presenting an artistic metaphor. These 3 videos better cover the multi-intent nature of the query and also keep the topical focus.

Query `presidential speeches quotes` follows a mono-intent response pattern, which is reflected in the initial top-25 results: the list contains one prevailing intent, i.e., to experience. However, some videos have higher potential to satisfy the user's intent better than others. Since this query is a mono-intent query and clearly focuses on quotes from historical presidential speeches, our approach reranks the list to move the videos with higher potential for intent satisfaction upward in the optimized list (`RR`): Video 1 was initially ranked at position 3, but the video intent classifier assigned it higher abilities to better satisfy this intent than the videos initially ranked 1 and 2. Video 2 is the same as in the initial list and video 3 was introduced in the top-3, also focusing on a historical quote. The refined list is much more focused than the initial list and can be further optimized using intent-informed visual reranking: $\mathtt{RR} + \mathtt{VRR}$ introduces a third video on historical quotes in the top-3 based on the visual similarities of thumbnails belonging to the same intent class. This demonstrates that the selective intent-informed visual reranking step leverages topic-specific visual similarities between videos with a given dominant intent and sharpens the intent focus at the top of the list.

VIII. Conclusion and Outlook

We have presented a novel approach that optimizes video search results lists for user intent by exploiting both textual and visual features extracted from initial lists. We performed a qualitative analysis which reveals that the engine produces intent responses following specific patterns. Our approach optimizes the initial ranking in light of these patterns and in a way that it concentrates the videos having the highest potential for intent satisfaction at the top of the results list. We evaluated our approach via extensive system-oriented evaluation and a large crowdsourcing user study, demonstrating that it results in substantially better information need satisfaction.

While this paper focused on revealing the potential of detecting and exploiting intent response patterns for intent-aware video search result optimization, our future work will focus on exploiting this potential towards maximizing the target result. This will involve studies on more sophisticated results list-reranking and -aggregation techniques and their effectiveness in this problem context. In this work we focused on intent optimization with respect to individual queries. An interesting direction for future work is to investigate intent-aware search over multiple queries, i.e., at the search session level. Since visual thumbnails have proven promise for further optimization, we aim to automatically extract the visual thumbnails from videos which are best suited for intent-aware visual reranking and to apply semantically richer representations like semantic visual concepts as visual content descriptors. We would like to conclude by pointing out that the need for intent-aware video retrieval will become increasingly pressing as the amount of video online continues to grow. For any given topic, an increasing number of videos on that topic will be available. If search engines are to provide users with highly relevant results, they must be able to satisfying not only the topical component of the query, but also the intent.

References

[1] L. Yang and A. Hanjalic, "Supervised reranking for web image search," in *Proc. ACM Int. Conf. Multimedia, ser. MM '10*, 2010, pp. 183–192.

[2] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua, "Bayesian video search reranking," in *Proc. 16th ACM Int. Conf. Multimedia, ser. MM '08*, 2008, pp. 131–140.

[3] Z.-J. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, T.-S. Chua, and X.-S. Hua, "Visual query suggestion: Towards capturing user intent in internet image search," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 6, no. 3, pp. 13:1–13:19, 2010.

[4] A. Hanjalic, C. Kofler, and M. Larson, "Intent and its discontents: The user at the wheel of the online video search engine," in *Proc. 20th ACM Int. Conf. Multimedia, ser. MM '12*, 2012, pp. 1239–1248.

[5] P. Chandar and B. Carterette, "Analysis of various evaluation measures for diversity," in *Proc. DDR Workshop European Conf. Information Retrieval, ser. DDR/ECIR '11*, 2011.

[6] S. Cronen-Townsend, Y. Zhou, and W. B. Croft, "Predicting query performance," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Research and Development in Information Retrieval, ser. SIGIR '02*, 2002, pp. 299–306.

[7] S. Wu and F. Crestani, "Data fusion with estimated weights," in *Proc. 11th ACM Int. Conf. Information and Knowledge Management, ser. CIKM '02*, 2002, pp. 648–651.

[8] T. Mei, Y. Rui, S. Li, and Q. Tian, "Multimedia search reranking: A literature survey," *ACM Comput. Surv.*, vol. 46, no. 3, pp. 38:1–38:38, Jan. 2014.

[9] Y.-H. Yang and W. Hsu, "Video search reranking via online ordinal reranking," in *Proc. IEEE Int. Conf. Multimedia and Expo, 2008*, 2008, pp. 285–288.

[10] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. 14th Annu. ACM Int. Conf. Multimedia, ser. MM '06*, 2006, pp. 35–44.

[11] W. H. Hsu, L. S. Kennedy, and Chang, "Video search reranking through random walk over document-level context graph," in *Proc. 15th ACM Int. Conf. Multimedia, ser. MM '07*, 2007, pp. 971–980.

[12] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *Proc. 2nd ACM Int. Conf. Web Search and Data Mining, ser. WSDM '09*, 2009, pp. 5–14.

[13] Y. Liu, T. Mei, and X.-S. Hua, "Crowdreranking: Exploring multiple search engines for visual search reranking," in *Proc. 32nd Int. ACM SIGIR Conf., ser. SIGIR '09*, 2009, pp. 500–507.

[14] R. Yan, J. Yang, and A. G. Hauptmann, "Learning query-class dependent weights in automatic video retrieval," in *Proc. 12th Annu. ACM Int. Conf. Multimedia, ser. MM '04*, 2004, pp. 548–555.

[15] L. S. Kennedy, A. P. Natsev, and S.-F. Chang, "Automatic discovery of query-class-dependent models for multimodal search," in *Proc. 13th Annu. ACM Int. Conf. Multimedia, ser. MM '05*, 2005, pp. 882–891.

[16] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. 2nd Int. Conf. Image and Video Retrieval, ser. CIVR'03*, Springer-Verlag, 2003, pp. 238–247.

[17] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li, "Learning to video search rerank via pseudo preference feedback," in *Proc. IEEE Int. Conf. Multimedia and Expo, 2008*, 2008, pp. 297–300.

[18] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.

[19] D. E. Rose and D. Levinson, "Understanding user goals in web search," in *Proc. 13th ACM Int. Conf. World Wide Web, ser. WWW '04*, 2004, pp. 13–19.

[20] B. J. Jansen, D. L. Booth, and A. Spink, "Determining the informational, navigational, and transactional intent of web queries," *Inf. Process. Manage.*, vol. 44, no. 3, pp. 1251–1266, May 2008.

[21] U. Lee, Z. Liu, and J. Cho, "Automatic identification of user goals in web search," in *Proc. 14th ACM Int. Conf. World Wide Web, ser. WWW '05*, 2005, pp. 391–400.

[22] H. Cao, D. H. Hu, D. Shen, D. Jiang, J.-T. Sun, E. Chen, and Q. Yang, "Context-aware query classification," in *Proc. 32nd Int. ACM SIGIR Conf. Research and Development in Information Retrieval, ser. SIGIR '09*, 2009, pp. 3–10.

[23] R. L. Santos, C. Macdonald, and I. Ounis, "Intent-aware search result diversification," in *Proc. 34th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, ser. SIGIR '11*, 2011, pp. 595–604.

[24] Y. Azar, I. Gamzu, and X. Yin, "Multiple intents re-ranking," in *Proc. 41st ACM Symp. Theory of Computing, ser. STOC '09*, 2009, pp. 669–678.

[25] B. J. Jansen, A. Spink, and J. Pedersen, "The effect of specialized multimedia collections on web searching," *J. Web Eng.*, vol. 3, no. 3, pp. 182–199, Dec. 2004.

[26] D. Tjondronegoro, A. Spink, and B. J. Jansen, "A study and comparison of multimedia Web searching: 1997-2006," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 9, pp. 1756–1768, Sep. 2009.

[27] C. Kofler and M. Lux, "Dynamic presentation adaptation based on user intent classification," in *Proc. 17th ACM Int. Conf. Multimedia, ser. MM '09*, 2009, pp. 1117–1118.

[28] J. Cui, F. Wen, and X. Tang, "Intentsearch: Interactive on-line image search re-ranking," in *Proc. 16th ACM Int. Conf. Multimedia, ser. MM '08*, 2008, pp. 997–998.

[29] H. Zhang, Z.-J. Zha, Y. Yang, S. Yan, Y. Gao, and T.-S. Chua, "Attribute-augmented semantic hierarchy: Towards bridging semantic gap and intention gap in image retrieval," in *Proc. 21st ACM Int. Conf. Multimedia, ser. MM '13*, 2013, pp. 33–42.

[30] T. Mei, X.-S. Hua, H.-Q. Zhou, and S. Li, "Modeling and mining of users' capture intention for home videos," *IEEE Trans. Multimedia*, vol. 9, no. 1, pp. 66–77, Jan. 2007.

[31] O. Alonso, D. E. Rose, and B. Stewart, "Crowdsourcing for relevance evaluation," *SIGIR Forum*, vol. 42, no. 2, pp. 9–15, Nov. 2008.

[32] C. Eickhoff, W. Li, and A. Vries, "Exploiting user comments for audio-visual content indexing and retrieval," in *Advances in Information Retrieval*, ser. Lecture Notes in Computer Science. Berlin, Germany: Springer, 2013, vol. 7814, pp. 38–49.

[33] O. Alonso and S. Mizzaro, "Can we get rid of TREC assessors? using mechanical turk for relevance assessment," ser. SIGIR '09 Works. on the Future of IR Eval.

[34] R. Blanco, H. Halpin, D. M. Herzig, P. Mika, J. Pound, H. S. Thompson, and T. Tran Duc, "Repeatable and reliable search system evaluation using crowdsourcing," in *Proc. 34th Int. ACM SIGIR Conf. Research and Development in Information Retrieval, ser. SIGIR '11*, 2011, pp. 923–932.

[35] S. Schmiedeke, P. Xu, I. Ferrané, M. Eskevich, C. Kofler, M. A. Larson, Y. Estève, L. Lamel, G. J. F. Jones, and T. Sikora, "Blip10000: A social video dataset containing spug content for tagging and retrieval," in *Proc. 4th ACM Multimedia Systems Conf, ser. MMSys '13*, 2013, pp. 96–101.

[36] S. Fox, K. Karnawat, M. Mydland, S. Dumais, and T. White, "Evaluating implicit measures to improve web search," *ACM Trans. Inf. Syst.*, vol. 23, no. 2, pp. 147–168, Apr. 2005.

[37] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Computer Vision, 2003*, 2003, vol. 2, pp. 1470–1477.

[38] Y. Li, B. Geng, L. Yang, C. Xu, and W. Bian, "Query difficulty estimation for image retrieval," *Neurocomputing*, vol. 95, no. 0, pp. 48–53, 2012.

[39] C. Kofler, M. Larson, and A. Hanjalic, "To seek, perchance to fail: Expressions of user needs in internet video search," in *Proc. 33rd European Conf. Advances in Information Retrieval, ser. ECIR'11*, Springer-Verlag, 2011, pp. 611–616.

[40] M. Lux and S. A. Chatzichristofis, "Lire: Lucene image retrieval: An extensible java cbir library," in *Proc. 16th ACM Int. Conf. Multimedia, ser. MM '08*, 2008, pp. 1085–1088.

[41] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Found. Trends Inf. Retr.*, vol. 2, no. 4, pp. 215–322, Apr. 2009.

[42] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychol. Rev.*, vol. 63, no. 2, pp. 81–97, Mar. 1956.

Christoph Kofler holds an M.Sc. and B.Sc. degree in Computer Science from Klagenfurt University, Austria and is currently pursuing his Ph.D. degree at Delft University of Technology, The Netherlands. His research interests include the field of multimedia information retrieval with focus on video search intent inference and its impact on search optimization. He is the recipient of the Google Doctoral Fellowship and the ACM MM Grand Challenge Award. He has held positions at Microsoft Research, China, Columbia University, USA and Google, USA.

Martha Larson holds an MA and Ph.D. in theoretical linguistics from Cornell University and a B.S. in Mathematics from the University of Wisconsin. Her research interest and expertise lie in the area of speech-and language-based techniques for multimedia information retrieval. She is co-founder of the MediaEval Multimedia Benchmark and has served as organizer of a number of workshops in the areas of spoken content retrieval and crowdsourcing. She has authored or co-authored over 100 publications. Currently, Dr. Larson is assistant professor in the Multimedia Information Retrieval Lab at Delft University of Technology. Before coming to Delft, she researched and lectured in the area of audio-visual retrieval at Fraunhofer IAIS and at the University of Amsterdam.

Alan Hanjalic is a professor of computer science, and head of the Multimedia Computing Group at the Delft University of Technology, The Netherlands. His research focus is on multimedia information retrieval and recommender systems. Prof. Hanjalic has been a member of the IEEE Technical Committee on Multimedia Signal Processing and the Steering Committee of the IEEE TRANSACTIONS ON MULTIMEDIA. He is Associate Editor-in-Chief of the IEEE MultiMedia Magazine and member of editorial boards of several scientific journals, including the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON AFFECTIVE COMPUTING and the International Journal of Multimedia Information Retrieval. He was General or Program Chair of major venues in the multimedia field, among which the ACM MM, ACM CIVR, ACM ICMR and IEEE ICME conferences.