Dealing with ambiguous Queries in Multimodal Video Retrieval

Luca Rossetto, Claudiu Tănase, and Heiko Schuldt

Databases and Information Systems Research Group
Department of Mathematics and Computer Science
University of Basel, Switzerland
{luca.rossetto|c.tanase|heiko.schuldt}@unibas.ch

Abstract. Dealing with ambiguous queries is an important challenge in information retrieval (IR). While this problem is well understood in text retrieval, this is not the case in video retrieval, especially when multimodal queries have to be considered as for instance in Query-by-Example or Query-by-Sketch. Systems supporting such query types usually consider dedicated features for the different modalities. This can be intrinsic object features like color, edge, or texture for the visual modality or motion for the kinesthetic modality. Sketch-based queries are naturally inclined to be ambiguous as they lack specification in some information channels. In this case, the IR system has to deal with the lack of information in a query, as it cannot deduce whether this information should be absent in the result or whether it has simply not been specified, and needs to properly select the features to be considered. In this paper, we present an approach that deals with such ambiguous queries in sketch-based multimodal video retrieval. This approach anticipates the intent(s) of a user based on the information specified in a query and accordingly selects the features to be considered for query execution. We have evaluated our approach based on Cineast, a sketch-based video retrieval system. The evaluation results show that disregarding certain features based on the anticipated query intent(s) can lead to an increase in retrieval quality of more than 25% over a generic query execution strategy.

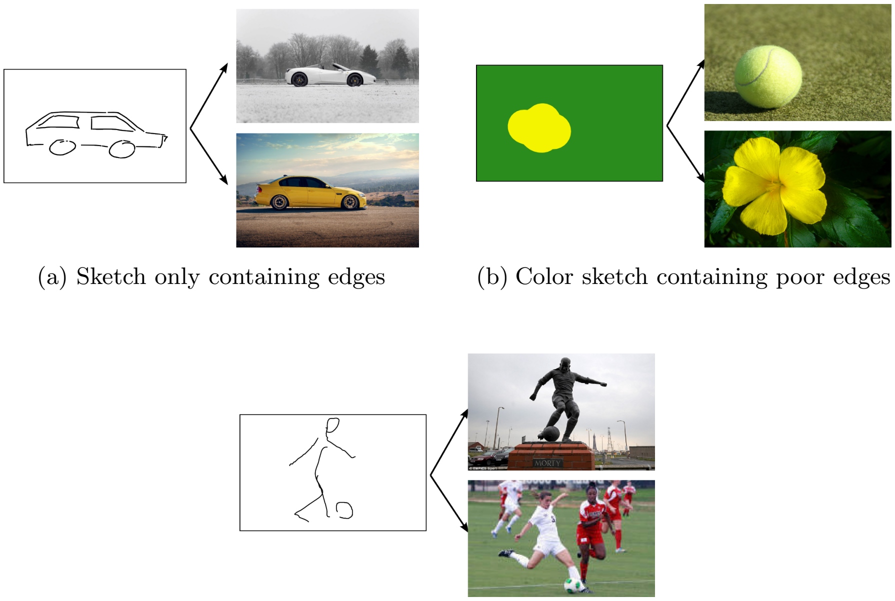
1 Introduction

Ambiguity is a property of most kinds of expression. While the problem of query ambiguity in text-based queries is well known, it differs from the situation found when dealing with multimodal queries. In text retrieval the difficulties lie primarily in the context-sensitivity of certain terms. Multimodal queries in contrast – such as the ones produced by query paradigms like Query-by-Sketch (QbS) or Query-by-Example (QbE) – additionally can have different meaning depending on the interpretation of the presence or absence or even the amount of information contained within each individual mode of expression. This is especially true in a QbS scenario where based on the query alone it is not possible to determine if a user did intend for a certain modality to contain no or limited information because the desired result should also lack this sort of content — or if the user was just not able to provide the information due to the lack of appropriate input devices, lack of artistic skills, or if he simply could not remember this aspect of the piece of content in question.

This problem is particularly relevant in video retrieval as video intrinsically comes with multiple *modalities* (e.g., visual, aural, text, kinesthetic) that are hardly provided jointly in a single query. Moreover, several of these modalities have different *information channels*, such as color, edge, or texture in the visual modality, or pitch, volume, speech, and music in the aural modality. Retrieval is based on inherent features characterizing these information channels individually; this can be done either by a single feature per information channel, or by the combination of several different features for one information channel (e.g., color histograms and color moments for the color information channel). For retrieval purposes, the lack of specification of information for one or several of the information channels or, even worse, the lack of specification for an entire modality has immediate consequences on the required selection and combination of features and thus on the retrieval results. If the wrong set of features is considered (e.g., features for which no or only poor information is given in the corresponding information channel of the modality), the query result is negatively impacted as these features might not all have the same selectivity. While the (amount of) information present in each channel is an inherent property of the content, the absence of information in a channel of the query could be due to the lack of care, memory or artistic skill of the user or because of unsuited means of input. It is therefore not possible to exactly determine the user's search intent based on the query and the amount of information present within its different channels. Hence, the absence of some information channel does not mean that this channel (and the features describing this channel) also does not have to be considered for query execution.

Query-by-Sketch video retrieval focuses on the two modalities that can be provided by users by means of sketches: the visual modality and the kinesthetic modality. For the visual part, users may provide line/contour sketches, color sketches, or a combination thereof via a sketch canvas. The kinesthetic modality can be specified in the form of flow fields via the same canvass to represent motion across frames. Alternatively, other user interfaces as for instance gesture-based UIs [8] can be used. In this paper, we will focus on all information that can be specified via a single sketch canvass (color, edges, and flow fields).

Figure 1 illustrates the problem of ambiguous queries resulting from sketch-based video queries. In Figure 1a, the query only contains a line-sketch of a car on a white background. It is therefore not clear whether the author of the query was explicitly looking for a white car on a white background or just did not care what color the car has. Similarly, Figure 1b contains color information, but without crisp edge information. Figure 1c lacks a specification of the kinesthetic modality (motion) in the sketched query. In this case, it is not obvious whether the query asks for a non-moving object such as the statue, or a video of a soccer scene in which motion plays an important role.



(a) Sketch only containing edges

(b) Color sketch containing poor edges

(c) Sketch without specified motion

Fig. 1: Examples of ambiguous sketch-based video queries

In this paper, we present an approach to disambiguate sketch-based queries in video retrieval, thereby solving the problem that inherently comes with this type of access where not all modalities and/or not all information channels within a modality are specified in a query. This is based on the consideration of the most likely user intent(s) that are anticipated by considering the modalities and information channels on which information is provided (or unintentionally absent) in the sketch and the combination of the retrieval results for these intents. We have evaluated the disambiguation strategy on top of Cineast, a novel multimodal video retrieval engine. The evaluation results show an increase in retrieval quality of more than 25%, measured in terms of the inverted retrieval rank, over a generic baseline approach which anticipates all possible query intents, independent of the information provided in the query sketch.

The contribution of the paper is twofold: First, we provide a detailed analysis of the problem of ambiguous queries in multimodal (sketch-based) video retrieval, especially with regard to the absence of certain modalities in the query (visual or motion), or the absence of information channels within a modality (e.g., colors or edges in the visual modality). Second, we present an approach to anticipate the possible query intent(s) of a user based on the information provided in the sketch or unintentionally left out to overcome these limitations in ill-posed queries. The evaluation results show that the consideration of only a subset of the available features, tailored to the user intent(s), leads to a significant improvement of the retrieval results compared with a generic combination of features, independent of the query content.

The remainder of the paper is organized as follows: Section 2 discusses the problems of ambiguity in multimodal video retrieval. In Section 3, we introduce the Cineast multimodal video retrieval system which is used as basis for the evaluation provided in Section 4. Section 5 surveys related work and Section 6 concludes.

2 Dealing with ambiguous Queries

In this section, we describe the method we use to deal with ambiguous multimodal queries in general and our concrete case with sketch-based video retrieval in particular. In our case, queries typically contain two modalities, the visual modality with the actual sketch and the motion modality represented by a flow field. The motion modality contains one information channel while the edge- and color information of the visual modality are treated differently, therefore resulting in two information channels for this modality.

2.1 Considering multiple intents

If the search intent of a user was known, a retrieval system could usually be optimized to best serve it by for example changing the individual parts used for measuring similarity or by adjusting their influence during the compilation of the final result list. Since determining the search intent for a general query on an unrestricted dataset is generally not reliably doable, we propose a different strategy. Even though one can usually not say with certainty what a user meant by a particular query, it is in most cases feasible to enumerate all possible intents and, depending on the dataset, even order them by expected probability. During query evaluation, a system can then perform different instances of the same similarity search optimized for different query intents and combine the individual result lists accordingly.

In our case with three information channels in the context of (sketch-based) video retrieval, we first have to analyze the query for its colorfulness, its crispiness (i.e., the amount of edge information contained within the query image) and its amount of movement. This analysis is done as a binary decision but could also be done in a more fine-gained way. The result of this analysis returns the possible query intent(s) of a user, called anticipated intent(s). These anticipated intent(s) are ordered by their expected probability, determined empirically by the analysis of actual queries, as shown in the second column in Table 1 (with the anticipated intents ordered with descending probability per cell). Each of these intents leads to one or multiple combinations of feature groups (again ordered per cell with decreasing probability in the rightmost column) which produces appropriate results for a given query.

Table 1: Anticipated intents for different query types, leading to different combinations of feature groups

Query Type			Anticipated Intent(s)	Combination of Feature Groups
colorful	crisp	moving	colorful crisp moving shot	(color, edge, motion) (color, edge) (color) (edge)
		still	colorful crisp still shot colorful crisp moving shot	(color, edge) (color, edge, motion) (color) (edge)
	blobby	moving	colorful blobby moving shot colorful crisp moving shot	(color, motion) (color, edge, motion) (color) (edge)
		still	colorful blobby still shot colorful crisp still shot colorful blobby moving shot colorful crisp moving shot	(color) (color, edge) (color, edge, motion)
colorless	crisp	moving	colorless crisp moving shot colorful crisp moving shot	(color, edge, motion) (edge, motion) (color, edge) (edge)
		still	colorless crisp still shot colorful crisp still shot colorless crisp moving shot colorful crisp moving shot	(edge) (color, edge) (edge, motion) (color, edge, motion)
	blobby	moving	colorless blobby moving shot colorful blobby moving shot colorless crisp moving shot colorful crisp moving shot	(motion)

2.2 Combining intents

Compared to the overall retrieval time for multiple different similarity measures, the duration of the result combination is usually negligible. For combination strategies which allow sub-queries to be performed in parallel, which is the case for late-fusion approaches, multiple combinations can be jointly considered without significant extra costs, because the sub-queries were performed without interdependency resulting in valid input for any number of combinations. In our case, the used combinations are listed in Table 1.

In what follows, we present three different ways, illustrated in Figure 2, in which this last combination step can be performed, together with a generic baseline approach we have used for the evaluation. The only commonality of the three proposed approaches is that duplicates are removed during combination.

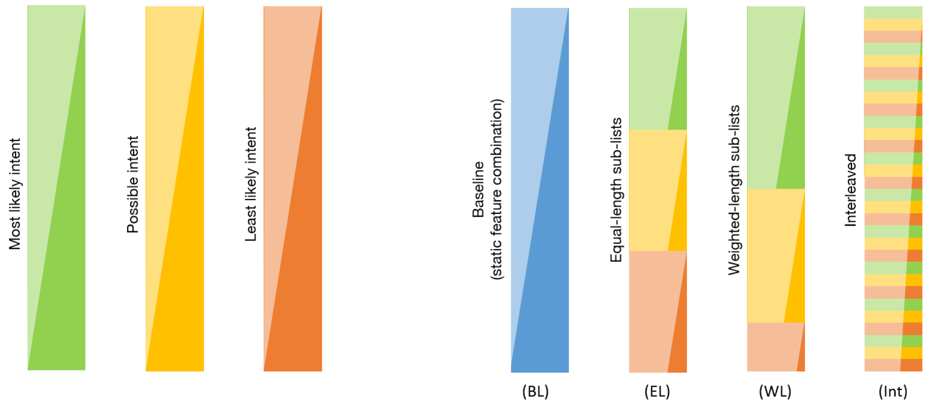


Fig. 2: Illustration of the different combination methods

Equal-length sub-lists (EL) In this combination scheme, each result list for the different anticipated intents gets an equal portion of the overall result list. The order of the sub-list is pre-determined based on which intent seems most likely given the query properties.

Weighted-length sub-lists (WL) In contrast to the previous evaluation scheme, the final list is comprised of the individual list in a ratio of $1 : 2 : 3$ and so on with the most likely query intent getting the largest part and the least likely intent getting the smallest. The ordering is the same as in the previous case.

Interleaved (Int) Other than in the previous two approaches, this combination method does not preserve the order of the results within the individual lists but rather interleaves the lists. In the Int approach, similar to EL, all lists contribute the same number of elements to the final result.

Baseline (BL) The easiest way to combine results from the individual features is to do so in a generic way, independently of the query. In this case it is beneficial to optimize the combination weights for the most common query type and use them for all queries.

3 Cineast

Cineast [13, 12] is a content-based video retrieval engine focused on Query-by-Sketch. It uses many different features in parallel to perform retrieval on different information channels such as color, edge, motion and text (in the form of subtitles if available) as well as meta-information. The results from these modules are combined using a score-based late fusion approach to construct a final result

list. Due to its modular architecture, feature modules can be added or removed easily enabling Cineast to use as many information channels as possible given a certain data set. Cineast also supports Query-by-Example as well as relevance feedback which enable the user to expand an initial result set in a promising direction towards the desired result.
Cineast groups individual features into groups which correspond to information channels. The relevant groups for this work are color, edge and motion. The following provides an overview of the features used in each group.

- Color: The color feature group contains low-level color representations such as global and local histograms, local statistical moments of color distribution as well as standardised descriptors such as the Color Layout Descriptor.
- Edge: The edge features consist of directional- as well as non-directional edge histograms with different spatial resolution.
- Motion: The motion features consist of local normalized directional motion histograms and intensity measures with different spatial resolutions.

4 Evaluation

We evaluate our approach on the OSVC1 dataset [14] consisting of 200 creative-commons web-videos with a wide range of content.

4.1 Evaluation procedure

To evaluate this approach, $N = 100$ query shots q out of the roughly 30'000 shots of the evaluation collection were randomly selected. From each of these shots, seven queries were created containing different combinations of information. Our performance metric is the Mean Inverted Rank (MIR), computed as the mean of the inverse rank value of the ground truth document across all queries
$MIR = \frac{1}{N}\left(\sum_q \frac{1}{R_q}\right)$.

In a late score fusion scheme, the relevance score of a retrieved document d is estimated to a weighted sum of relevance scores from individual feature estimators $score(d) = \sum_i w_i \cdot score(d, i)$. In order to learn optimal weights w_i we apply an optimization step of the weight vector w_i with the MIR as objective function, using the Sequential Least Squares Quadratic Programming method. In our experiments, the MIR has been optimized to a mean value of $MIR = 0.9288$ across 10 folds, after less than 8 iterations in each fold.

Weights are not changed across combination methods for the feature groups that are considered in a query.

In our case, a query has three information channels, color, edge, and motion, each of which can be omitted based on the user's intent or ability of expression. This leads to eight possible combinations if the decision whether or not to consider a channel of information is modeled as a binary one as we do in this evaluation. One of the eight possible combinations contains no information at all and can therefore be omitted.

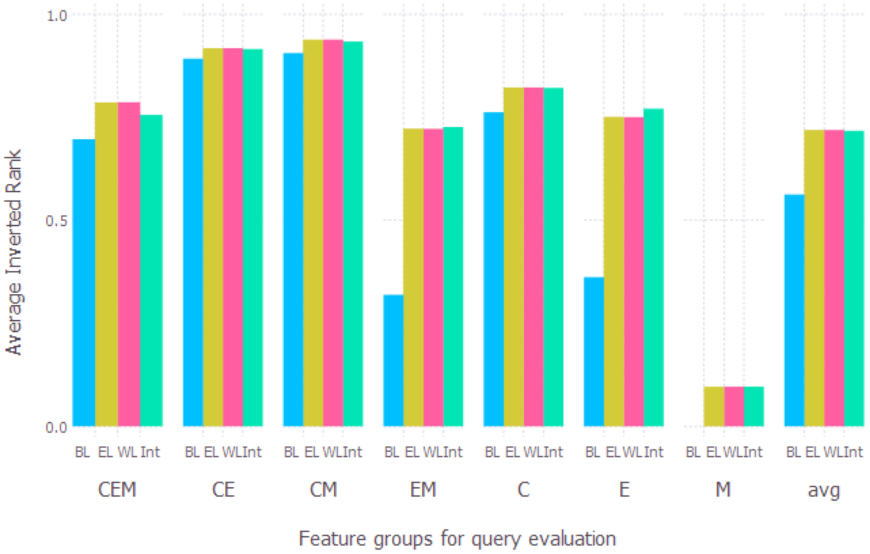


Fig. 3: Average inverted ranks for top 50 results

For each of the 100 shots we generate seven queries by omitting different information channels. To remove color information, the representing image is converted to monochrome and edge information is removed by low-pass filtering the image. Every information channel which is not artificially removed is designated with a character in the query name. A query where no information channel has been removed is designated as *CEM* (for containing *C*olor, *E*dge and *M*otion), while for example a query where the color information has been removed is designated as *EM*. It is important to note that due to the random selection, some of the shots do not contain certain information channels and therefore artificially removing them has no effect e.g., a grayscale shot will remain unchanged in *EM* mode. Similarly, removing the information channels present and keeping those originally without content results in empty or near-empty queries. For this evaluation, these combinations are not manually excluded to avoid the introduction of biases. As a consequence, ground truth is one unique result per query, making this a case of a known-item search.

4.2 Measurements

We compare the three described methods of combining multiple intents, EL, WL, and Int, against the baseline method BL which uses a static combination of weights set to produce generally good results. We measure for each of the seven variants of the 100 queries the number of correctly retrieved results as well as the average inverted rank for a result list with length 50. The number of results

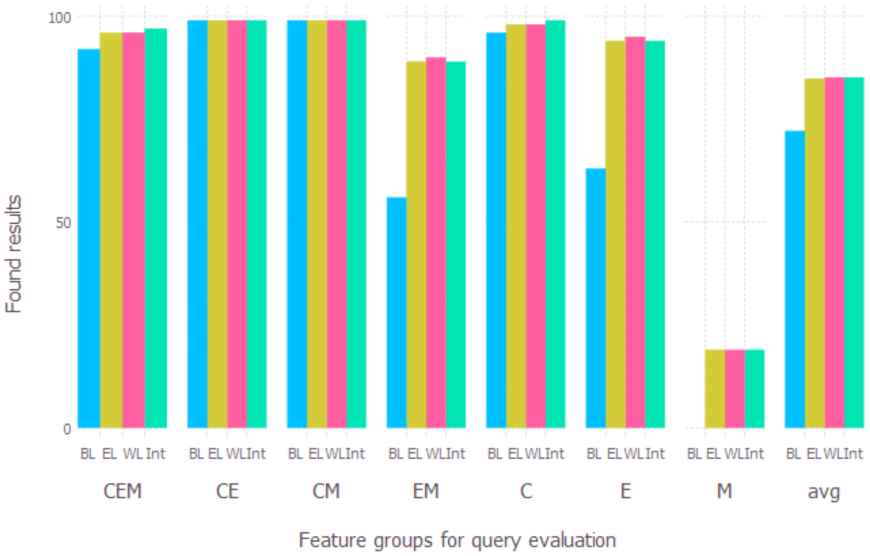


Fig. 4: Number of found correct results in the top 50

returned by the individual modules was always capped at twice the number of overall results.

4.3 Results

Figures 3 and 4 show the average inverted rank and number of found results out of the 100 possible per method and query type respectively. It can be seen in both figures that the BL method never outperforms any of the three proposed methods EL, WL, and Int. It can also be seen that EL, WL, and Int produce consistently high results for all but one query type. The comparatively bad results for the queries only containing motion is due to the low overall information contained in this channel, which also explains the fact that the BL method was unable to produce a single correct result for this query type. On average, it appears to make no difference if the result lists for the different possible intents are concatenated (EL and WL) or interleaved (Int), as long as they are considered at all. For some query types, some combination methods appear to perform slightly better than others but this can as well be due to artefacts in the query generation process or due to inherent properties of the used dataset. The inverted rank improves on average from 0.56 to 0.71 which is a relative improvement of 26.8% and the average number of correctly retrieved results increases from 72 to 85 out of 100. Perhaps counter-intuitively, results confirm that by intelligently selecting features, thus disregarding certain feature information, the retrieval performance in terms of both MIR and the number of correctly retrieved objects actually increases.

5 Related Work

While ambiguous queries in search engines are a commonly understood issue, most of the papers that have dealt with quantitative measurements or disambiguation strategies [3, 11, 16–18] have focused exclusively on web search/text retrieval. The disambiguation of text searches is achieved using external data such as ontologic information [11] which cannot be easily transferred to video search, as semantic video indexing is still an open problem in itself. A user-centric disambiguation method is presented in [19], where visual examples presented as the user formulates his query, help a content-based image retrieval system better understand the intent. Similarly, in [1] mutual relevance feedback from the user to the system and vice-versa contribute to refining the retrieval results in a multimodal context.

Ranking results is a key problem of information retrieval systems. In the traditional IR view, ranking is achieved by sorting results by their estimated relevance to the query. In later years, machine learning techniques have been used to learn better ranking models from training data. Techniques based on SVMs [2] and neural networks [6] have been applied to learn rankings of documents based on a corpus of queries with corresponding relevances manually rated or derived from external measurements (e.g., click-through rate for web search engine results [7]). The problem of optimally selecting a subset of a large or dynamically growing set of features has been studied by the IR literature as feature selection. There is a significant body of work dedicated to selecting features in order to improve ranking [5, 9]. These selectors however work in a query-independent way and thus cannot decide which features are relevant on a specific query.

Content-based video retrieval systems rely on multiple (typically in the tens) heterogeneous features to index the video content. Although more advanced methods exist [15], late score fusion is a simple yet effective multimodal feature fusion method widely used in practical systems because of its scalability to huge datasets, as shown in multimedia information retrieval challenges such as TRECVID [10] and MediaEval [4]. Essentially, late fusion takes estimated scores for a document and applies a set of precomputed weights.

6 Conclusion

In this paper, we have introduced an approach that is able to deal with ill-specified and thus ambiguous queries in sketch-based multimodal video retrieval. We focus on the visual and the kinesthetic modality, which are specified via sketches in a sketch canvas by means of edges, color, and flow fields. The absence of one or several of these information may be intentional and thus should be respected in the result list (e.g., no motion expected in result) – or it may simply be a result of lack of memory, artistic skills, or other reasons. In the latter case, the result should not be impacted by this lack of information. We have presented an approach that first anticipates the query intent(s) of a user based on the information specified in a query sketch. Based on the anticipated user intent(s), the necessary feature groups are chosen for query execution. In an evaluation with the sketch-based video retrieval engine Cineast, we have shown the effectiveness of this approach which significantly improves retrieval quality over a baseline approach agnostic to the query intent at negligeable extra computational cost.

While we have concentrated in this paper on the visual and kinesthetic modalities, the approach can be easily extended to other modalities and the information channels characterizing these modalities. In the IMOTION project, we are currently working on the combination of hand-drawn sketches and spoken queries which will add support for the aural modality. We will also evaluate further approaches to specify information on the motion modality beyond flow fields. In addition, we plan to analyze the impact of a more fine-grained differentiation of the modalities and information channels beyond a binary selection. Similarly, further differentiations within a channel will be considered. This will allow to apply individual weights for features and feature groups dependent on the amount and/or quality contained within an information channel.

Acknowlegements

This work was partly supported by the Swiss National Science Foundation in the context of the CHIST-ERA (European Coordinated Research on Long-term Challenges in Information and Communication Sciences & Technologies ERA-Net) project IMOTION (Intelligent Multi-Modal Augmented Video Motion Retrieval System), contract no. 20CH21_151571.

References

1. Arnon Amir, Marco Berg, and Haim Permuter. Mutual relevance feedback for multimodal query formulation in video retrieval. In *Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval*, pages 17–24. ACM, 2005.
2. Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96. ACM, 2005.
3. Steve Cronen-Townsend and W Bruce Croft. Quantifying query ambiguity. In *Proceedings of the second international conference on Human Language Technology Research*, pages 104–109. Morgan Kaufmann Publishers Inc., 2002.
4. Maria Eskevich, Robin Aly, David Racca, Roeland Ordelman, Shu Chen, and Gareth JF Jones. The search and hyperlinking task at mediaeval 2014. 2014.
5. Xiubo Geng, Tie-Yan Liu, Tao Qin, and Hang Li. Feature selection for ranking. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 407–414. ACM, 2007.

6. Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large margin rank boundaries for ordinal regression. *Advances in neural information processing systems*, pages 115–132, 1999.
7. Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 133–142. ACM, 2002.
8. Ihab Al Kabary and Heiko Schuldt. Using hand gestures for specifying motion queries in sketch-based video retrieval. In *Proceedings of the 36th European Conference on Information Retrieval (ECIR 2014), Research Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 733–736, Amsterdam, The Netherlands, April 2014. Springer.
9. Jasmina Novaković, Perica Štrbac, and Dušan Bulatović. Toward optimal feature selection using ranking methods and classification algorithms. *Yugoslav Journal of Operations Research ISSN: 0354-0243 EISSN: 2334-6043*, 21(1), 2011.
10. Paul Over, Georges Awad, Martial Michel, Johnatan Fiscus, Greg Sanders, Wessel Kraaij, Alan F Smeaton, and Georges Quénot. Trecvid 2014- an overview of the goals. *Tasks, Data, Evaluation Mechanisms, and Metrics In Proceedings of TRECVID*, 2014.
11. Guang Qiu, Kangmiao Liu, Jiajun Bu, Chun Chen, and Zhiming Kang. Quantify query ambiguity using odp metadata. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 697–698. ACM, 2007.
12. Luca Rossetto, Ivan Giangreco, Silvan Heller, Claudiu Tănase, and Heiko Schuldt. Searching in video collections using sketches and sample images – the cineast system. In *MultiMedia Modeling*. Springer, 2016.
13. Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. Cineast: A multi-feature sketch-based video retrieval engine. In *Multimedia (ISM), 2014 IEEE International Symposium on*, pages 18–23. IEEE, 2014.
14. Luca Rossetto, Ivan Giangreco, and Heiko Schuldt. OSVC – Open Short Video Collection 1.0. Technical report, University of Basel, 2015.
15. Cees GM Snoek, Marcel Worring, and Arnold WM Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 399–402. ACM, 2005.
16. Ruihua Song, Zhenxiao Luo, Ji-Rong Wen, Yong Yu, and Hsiao-Wuen Hon. Identifying ambiguous queries in web search. In *Proceedings of the 16th international conference on World Wide Web*, pages 1169–1170. ACM, 2007.
17. Nenad Stojanovic. On analysing query ambiguity for query refinement: The librarian agent approach. In *Conceptual Modeling-ER 2003*, pages 490–505. Springer, 2003.
18. Kilian Quirin Weinberger, Malcolm Slaney, and Roelof Van Zwol. Resolving tag ambiguity. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 111–120. ACM, 2008.
19. Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, Zengfu Wang, Tat-Seng Chua, and Xian-Sheng Hua. Visual query suggestion: Towards capturing user intent in internet image search. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 6(3):13, 2010.