2009 Seventh International Workshop on Content-Based Multimedia Indexing

Query Generation From Multiple Media Examples

Reede Ren, Joemon M. Jose
Computing Science Dept., University of Glasgow
17 Lilybank Gardens, Glasgow, UK, G12 8QQ
reede,jj@dcs.gla.ac.uk

Abstract

This paper exploits a media document representation called feature terms to generate a query from multiple media examples, e.g. images. A feature term denotes a continuous interval of a media feature dimension. This approach (1) helps feature accumulation from multiple examples; (2) enables the exploration of text-based retrieval models for multimedia retrieval. Three criteria, minimised χ^2, minimised AC/DC and maximised entropy, are proposed to optimise feature term selection. Two ranking functions, KL divergence and BM25, are used for relevance estimation. Experiments on Corel photo collection and TRECVid 2006 collection show the effectiveness in image/video retrieval.

1. Introduction

The employment of multiple query examples is a popular query scenario in multimedia information retrieval (MIR). There are two common scenes: (1) users submit several example images or video clips at the beginning of a query [1]; (2) retrieval systems gradually find new examples by relevance feedback and query expansion [2]. Yan *et al.* [3] assert that multiple query examples substantially reduce "word mismatch" and facilitate the formulation of a good query. Query generation from multiple examples is to seek a concept model (query) across a set of knowledge sources, *i.e.* query examples, and a research problem of information fusion, wich creates [3]. Note that media features are of difference similarity measurements, *e.g.* euclidian distance for RGB color and intersection distance for edge histogram. A high computational complexity is hinted in query generation and document relevance estimation.

The literature can be roughly categorised into two groups, early fusion and late fusion. Early fusion directly learns a query from extracted features. Various machine learning strategies have been proposed to decide on an optimal solution, *e.g.* active learning [2] and automatic labelling [3]. However, a query example set is usually too small to support a robust analysis [4]. This will seriously degrade retrieval performance if some relevant features are ignored and if some non-relevant features are over-weighted. In late fusion, query examples are submitted individually like a group of sub-queries. Many empirical ranking schemes are used to merge sub-query results [3]. However, this leads to an intensive tuning on model parameters with respect to a query [1]. It is difficult to extend late fusion approaches for general large scale MIR.

Query generation is a pattern mining not only from a query example set, but also in the context of a document collection. Zhai *et al.* [5] assert that the retrieval is a statistical decision based on the variance of term distributions in both document collection and a query. Normalising feature distribution in a collection enlarges the significancy of a query. Collection knowledge therefore alleviates the uncertainty caused by the small query example set. We propose a three-step query generation from multiple examples: (1) project media features into a vocabulary called *feature terms*; (2) define a statistical measurement on the distribution of *feature terms* to optimise collection description, as well as to create a unified query; and (3) employ text retrieval models to estimate relevance. The contributions are: (1) the usage of collection knowledge to facilitate query generation; (2) an efficient approach for query generation and collection representation, which accumulates characteristics from media documents, especially low-level features; (3) a mixed ranking scheme across medias and documents for relevance estimation, which avoids complex distance computation and parameter tuning.

The remainder of this paper is structured as follows. Section 2 surveys the literature related to term distribution in text retrieval and also term-like feature extraction in MIR. Section 3 justifies feature term selection by proposing three criteria, minimised χ^2, maximised entropy and minimised AC/DC. The retrieval system and relevance estimation are presented in Section 4. Three parts of experiments are addressed in Section 5: term selection, retrieval experiments on the Corel photo collection and the TRECVid 2006 video collection. A brief conclusion is found in Section 6.

2. Related Work

We define a feature term as a range interval of a feature dimension. As our approach exploits the same principles employed in statistical text retrieval, we begin with an discussion about text term distribution.

As an important aspect in term weighting, text term distribution has been well discussed for the justification

978-0-7695-3662-0/09 $25.00 © 2009 IEEE
DOI 10.1109/CBMI.2009.13

of retrieval models [6]. Harter *et al.* [7] propose that a term should follow a 2-Poisson distribution, because term appearance is Boolean and sparsely distributed. Margulis *et al.* [8] extend this model to N-Poisson. They argue that N-Poisson might have provided a more precise estimation than 2-Poisson does, if a term actually followed a Poisson-like distribution. Several class numbers from two to seven were evaluated on real document collections [8], but no specific class number of Poisson combined model shows a significant out-performance. Amati *et al.* [6] simulate a retrieval process by a Bernoulli distribution. Amati *et al.* suggest a uniform term distribution, as the joint probability of multiple terms is so small that a simple uniform distribution is good enough for the modelling of term distribution.

In MIR, visual words [9] or concepts [10] are term-like representation for media documents, although working for high-level rather than low-level features. This is partially because low-level features (1) are continuously distributed and (2) require complex similarity measurements for content description. A visual word is conceptually similar to a text word: the close association with semantics and the Boolean nature, *i.e.* present or absent in a document. However, it is difficult to employ such an approach to present/analyse a large collection of general multimedia data. This is partly because of (1) domain dependency among concepts [10], (2) the lack of common concept definitions [11] and (3) the sparse distribution of concepts [11]. These facts result in the ineffectiveness of traditional ranking schemes. Nevertheless, the imperfection in the technique of automatic annotation reduces the reliability of concepts in retrieval [10]. Low-level features hence are widely used for media document indexing.

In this work, we follow a uniform distribution for *feature term* extraction. This is because this hypothesis leads to a superior retrieval performance in [6]. Moreover, the computational cost of a uniform distribution is significantly lower than N-Poisson. This is essential in MIR.

3. Feature Term Extraction

In this section, we describe methods which identify *feature terms* from a collection. The extraction of a *feature term* is a projection from a multiple valued N-dimensional variable to an integer/class label or a boolean vector of integer appearance. For example, the classification of a RGB colour into four classes can be depicted as $[0, 255]^3 \rightarrow \{0, 1, 2, 3\}$ or $[0, 255]^3 \rightarrow \{0, 1\}^4$ for the appearance of class label 0,1,2 and 3. This is symbolised as a function $\hat{f} : [0, K]^N \rightarrow \{0, 1, \ldots, M-1\} \sim \{0, 1\}^M$, where K denotes variable range and M the number of integers. We regard these integers as *feature terms*. Since dimensions in a low-level feature are independent from each other, we take the one-dimensional case where $N = 1$. This means that we process every feature dimension individually.

For a collection D, *the frequency of a feature term* f_t is the times that document features fall into a range interval $t \in [0, M)$.

$$f_t = |D_t|, D_t = \{d|\hat{f}(d) = t, d \in D\} \tag{1}$$

where d is a document in D. The probability of a *feature term* t is,

$$p(t) = \frac{f_t}{\sum_{i=0}^{M-1} f_i} \tag{2}$$

3.1. Selection Criterion

Some statistical criteria are necessary to justify an optimal solution. Given the uniform assumption [6], we propose three criteria, minimised χ^2(chi-square) test, maximised entropy and minimised AC-DC rate,

χ^2 **Test** computes the similarity of a sample sequence from a given distribution. As the optimised term probability is $\hat{p}(t) = \frac{1}{M}$, χ^2 test is defined as follows.

$$\chi^2(M) = \sum_{i=0}^{M-1} \frac{(p(t_i) - \hat{p}(t_i))^2}{\hat{p}(t_i)} = \sum_{i=0}^{M-1} \frac{(Mp(t_i) - 1)^2}{M} \tag{3}$$

The criterion of minimised χ^2 test is,

$$I_{\chi^2} = \arg\min_M \chi^2(M) \tag{4}$$

Entropy measures information gain brought by a given term selection.

$$Entropy_s(M) = -\frac{1}{\sqrt{M-1}} \sum_{i=0}^{M-1} p(t_i) \log(p(t_i)) \tag{5}$$

A high entropy indicates a good selection.

$$I_{entropy} = \arg\max_M Entropy_s(M) \tag{6}$$

AC-DC rate computes the variance of a data sequence from the average. For a frequency sequence $f_0, f_1, \ldots, f_{M-1}$, the DC parameter (Equation 7) denotes the mean while the first AC parameter (Equation 8) refers to the strongest deviation.

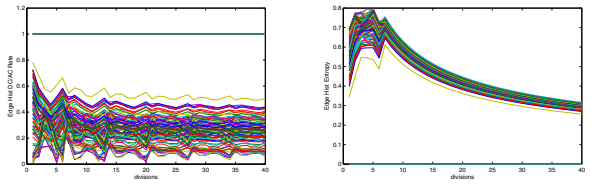
$$DC = \frac{1}{M} \sum_{n=0}^{M-1} f_n \tag{7}$$

$$AC = \frac{1}{M} \| \sum_{n=0}^{M-1} f_n e^{-\frac{2\pi i n}{M}} \| \tag{8}$$

The rate of AC-DC (Equation 9) reflects the bias of the frequency sequence away from the average. A low $R_{AC/DC}$ is preferred.

$$R_{AC/DC} = \frac{AC}{DC} \sim \sum_{n=0}^{M-1} f_n e^{-\frac{2\pi i n}{M}} \tag{9}$$

Figure 1 displays criterion value distribution (y-axis) with



(a) ACDC Rate (b) Entropy

Figure 1. Criterion value distribution for term selection in 80-dim edge histogram

different feature term selections (x-axis) for 80-dim edge histogram in the TRECVid 2006 collection. Favoured maximum/minimums appear on all dimensions, which indicates the effectiveness of respective criterion.

4. Collection Representation and Retrieval System

In this section, we describe the steps in feature terms based MIR. The system framework is shown in Figure 2. Four MPEG-7 low-level features are extracted, including colour layout (12 dims), dominant colour (7 dims), edge histogram (80 dims) and homogeneous texture (53 dims). A boolean vector of feature terms is computed to represent a media document while a frequency vector stands for a collection. The number of feature terms is decided by the criteria that are outlined in Section 3.1.

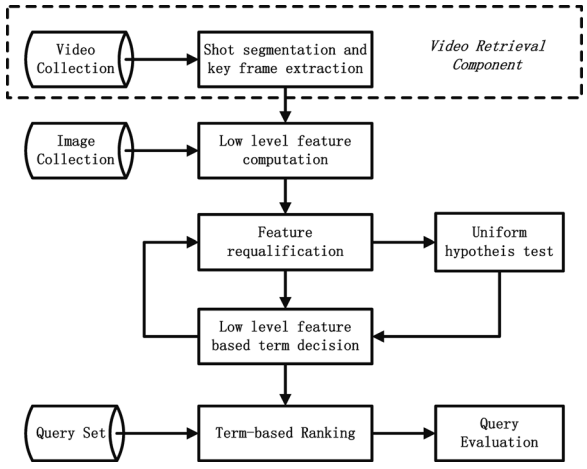


Figure 2. Retrieval System Framework

4.1. Document Representation

Let $V = \{t_1, t_2, \ldots, t_n\}$ be the vocabulary of feature terms. A media document d is therefore presented by a Boolean vector based on the vocabulary.

$$I_{d,V} = \{I_{t_1,d}, \ldots, I_{t_n,d}\} \tag{10}$$

where $I_{t,d} = 1$,iff $t \in d$, otherwise $I_{t,d} = 0$. A query Q is described by a frequency vector of feature terms which accumulates the appearances of feature terms in all examples q.

$$C_{Q,V} = \sum_{q \in Q} I_{q,V} \tag{11}$$

This defines a vector representation of feature terms for document and query. We use text retrieval ranking functions for relevance estimation.

4.2. KL Divergence Ranking

The negative KL divergence (Equation 12) [6] compares term distribution bias between a query Q and a media document d.

$$\begin{aligned} -D_{K,L}(\theta_Q|\theta_d) &= H(\theta_Q) - H(\theta_Q, \theta_d) \\ &= H(\theta_Q) + \sum_{t \in V} \theta_{t,Q} \log \theta_{t,d} \end{aligned}$$

where H is the entropy, t denotes a term in the vocabulary V, θ_Q and θ_d stand for the representation for the query and a document, respectively. $\theta_{t,Q}$ and $\theta_{t,d}$ are shorthand for $P(t|\theta_Q)$ and $P(t|\theta_d)$. Note that $H(\theta_Q)$ is constant for a given query,

$$-D_{K,L}(\theta_Q|\theta_d) \sim \sum_{t \in V} \theta_{t,Q} \log \theta_{t,d} \tag{12}$$

Since the appearance of a feature term $I_{t,d}$ is Boolean, the relevance status value is defined as follows.

$$RSV(d;Q) = \sum_{t \in V} \theta_{t,Q} \log \theta_{t,d} I_{t,d} \tag{13}$$

4.3. BM25 Ranking

We also propose an approach based on the BM25 model [6]. For a media document, the frequency of a feature term t in document d, $f_{t,d}$ is the binary $I_{t,d}$. Note that we rely on images or keyframes from video shots for retrieval. Unlike text documents, images, especially keyframes from a video, are of constant size. Therefore, adjustments on term frequency, *e.g.* the normalisation of document size, is unnecessary. The relevance status value is computed as,

$$RSV(d;Q) = \sum_{t \in V} IDF(t) I_{t,d} C(t,Q) \tag{14}$$

where $C(t,Q)$ is the appearance frequency of a feature term t in a query Q (Equation 11) and $IDF(t)$ is similar to inverse document frequency (Equation 15).

$$IDF(t) = \log \frac{N - n(t) + 0.5}{n(t) + 0.5} \tag{15}$$

where N is the document number in a collection and $n(t)$ is the number of documents with a given feature term t.

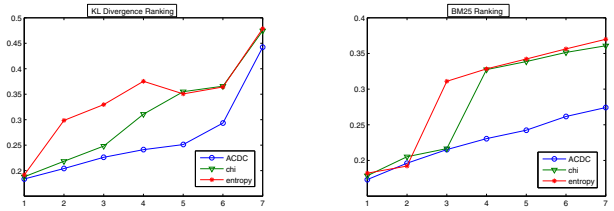
5. Experiment

The evaluation collection involves the Corel photo collection and the TRECVid 2006 video collection for the effectiveness of feature terms in image and video retrieval.

5.1. Corel Collection

We randomly chose 50 categories from the Corel photo collection in which each category contained 100 images. An experimental collection of 5,000 images is therefore created. Seven images were randomly selected from every category as query examples, which were gradually submitted to simulate a query with 1 to 7 examples. The top 100 relevant images were returned as retrieval results. This process was repeated five times. In addition, the precision and recall were equal here.

We compare the average precision (y-axis) for one to seven query examples (x-axis) with different criteria of term selections and for different ranking schemes (see Figure 3 for edge histogram). The following conclusions have been reached: (1) more query examples improve the retrieval performance of feature term based approaches; (2) maximised entropy is the best choice for *feature term selection* in a small document collection such as the Corel; and (3) KL ranking out-performs the BM25 ranking.



(a) KL Ranking (b) BM25 Ranking

Figure 3. Retrieval Performance of EdgeHist

Table 1 lists the precision achieved by different feature or feature combinations under the KL and BM25 ranking schemes. The combination of colour layout and edge histogram shows the best retrieval performance over all. In summary, these experiments in the Corel Photo collection prove that feature-term based approaches are effective and robust in image retrieval.

5.2. TRECVid Collection

TRECVid 2006 collection includes 24 content-based queries (Topic 173-196), such as *"find shots of Condoleeza Rice (Topic 194)"*. Each query is presented by between seven to eleven image examples and other annotations, *e.g.*text tags and audio clips. However, low-level features are ineffective for most queries [12]. Natsev *et al.* [11] regard low-level features as an additional knowledge source and argue that little improvement could be achieved by low-level features comparing with text and high-level concepts. To avoid bias, we take two low-level feature based retrieval methods in early TRECVid workshops [13] as baseline, including direct comparison and KNN clustering. Direct comparison computes a mixed Euclidean distance to identify the closet or most similar keyframes to a query. The kNN clustering groups keyframes into K clusters, each of which contains 600 keyframes. The top two closest clusters are returned as results. Returned documents are re-ranked by visual similarity.

Table 2 lists the performance of dominant colour by num-rel-ret (number of relevant documents in the top 1000 returned documents) [13]. A high num-rel-ret denotes a good performance. Feature terms collected by minimised ACDC achieve the best. Feature term based approaches (1)outperform kNN, (2)are comparable with direct comparison in performance, but require a low computational cost. More results are shown in Figures 4, 5 and 6 for colour layout, edge histogram and homogeneous texture, respectively.

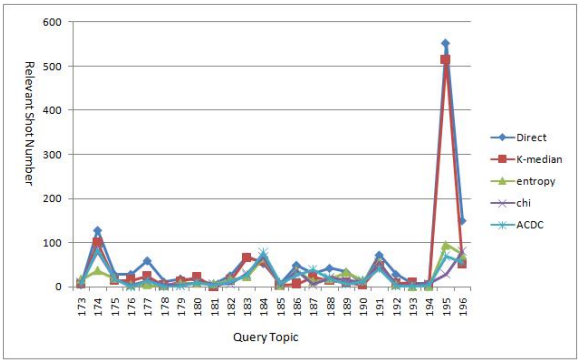


Figure 4. Num-Rel-Ret by Color Layout

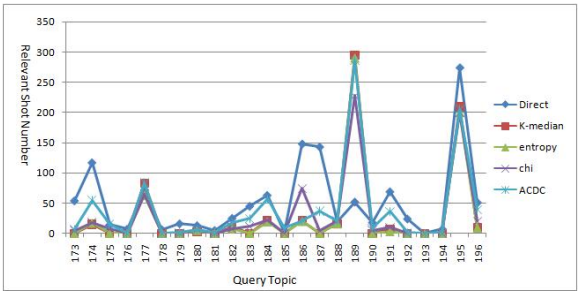


Figure 5. Num-Rel-Ret by Edge Histogram

We also compute average-precision (AP) over all topics. Figure 7 shows the AP of colour layout. Max-Max denotes

Feature	Ranking Function	Precision at Different Example Set Size						
		1	2	3	4	5	6	7
Dominant Colour	KL	*0.172*	*0.265*	*0.286*	*0.381*	*0.382*	*0.384*	**0.387**
	BM25	0.036	0.138	0.136	0.135	0.135	0.150	0.150
Edge Histogram	KL	*0.192*	*0.299*	*0.330*	*0.375*	*0.351*	*0.364*	**0.479**
	BM25	0.182	0.192	0.311	0.328	0.342	0.356	0.370
Homogeneous Texture	KL	*0.179*	*0.201*	*0.219*	*0.230*	*0.242*	*0.234*	**0.260**
	BM25	0.167	0.192	0.207	0.218	0.226	0.233	0.247
Colour Layout	KL	*0.203*	*0.330*	*0.369*	*0.392*	*0.413*	*0.432*	**0.548**
	BM25	0.142	0.142	0.238	0.234	0.232	0.233	0.231
Colour Layout & Edge Histogram	KL	*0.240*	*0.332*	0.451	0.488	*0.681*	*0.722*	0.730
	BM25	0.227	0.294	*0.485*	*0.502*	0.654	0.718	**0.736**
All Features	KL	0.255	0.184	0.211	0.350	0.440	*0.580*	0.694
	BM25	*0.262*	*0.200*	*0.277*	*0.411*	*0.453*	0.555	**0.703**

Table 1. Average Precision/Recall Under Maximised Entropy Criterion

Topic	Direct	kNN	Entropy	χ^2	ACDC	Topic	Direct	KNN	Entropy	χ^2	ACDC
173	5	1	11	12	*13*	**174**	43	29	34	45	*51*
175	18	5	21	16	*24*	**176**	*7*	3	*7*	*7*	*7*
177	23	3	*25*	*25*	7	**178**	1	1	8	8	*13*
179	*6*	0	2	4	2	**180**	0	1	9	*11*	2
181	*8*	0	1	7	4	**182**	*25*	2	8	15	17
183	*33*	7	11	21	22	**184**	30	6	45	46	*51*
185	8	1	3	*10*	*10*	**186**	71	12	56	*79*	57
187	24	2	12	28	*50*	**188**	25	2	9	*38*	29
189	24	0	63	54	*74*	**190**	3	1	7	11	*11*
191	49	5	63	*76*	75	**192**	2	8	*10*	9	1
193	2	0	2	8	*9*	**194**	5	0	6	8	*9*
195	87	0	59	95	*101*	**196**	*58*	26	20	49	46
average for all	23.21	4.79	20.50	28.42	28.54	-	-	-	-	-	-

Table 2. Num-Rel-Ret of dominant colour

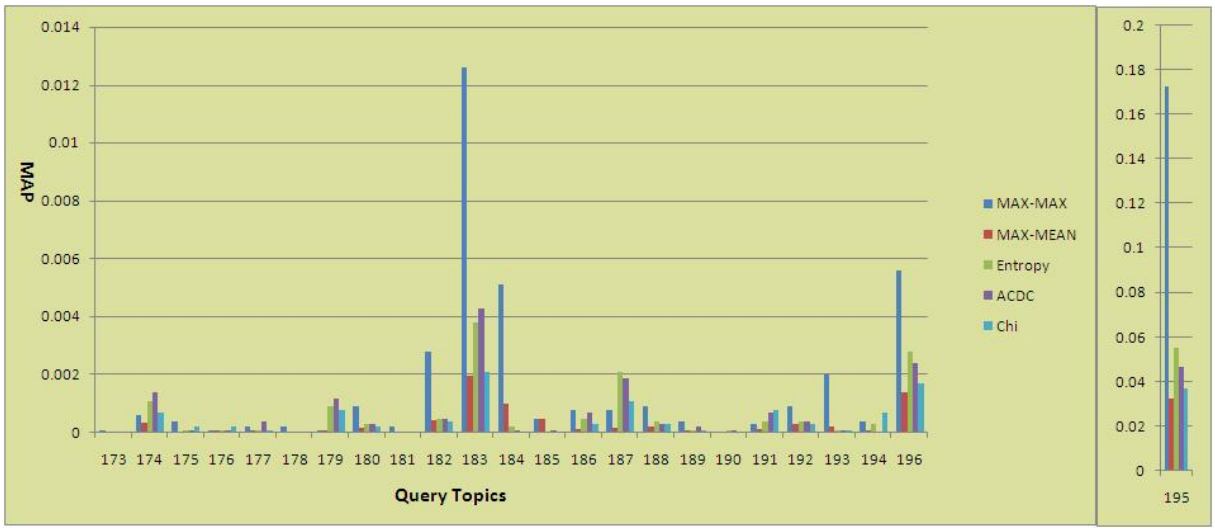


Figure 7. Average Precision of Colour Layout Query

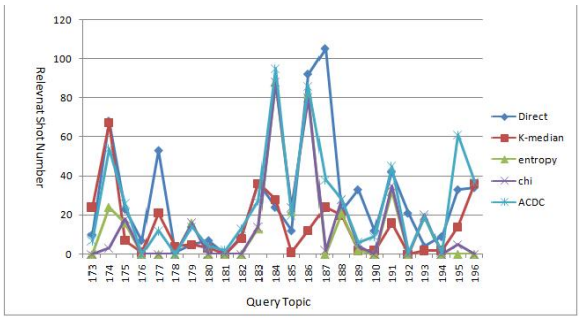


Figure 6. Num-Rel-Ret by Homogeneous Texture

the AP of an oracle that collects all relevant documents found by individual examples and direct comparison [12]. Max-Mean is the average AP achieved by individual examples. Max-Max is the performance upper boundary of late fusion approaches and Max-Mean refers to the baseline. The difference between Max-Max and Max-Mean indicates the number of examples which contribute positive knowledge or effective for a query. In most topics, the AP of feature term approaches are below MAX-MAX but above MAX-Mean. This proves the effectiveness of feature terms. In Topic 174,176,177,179,187 and 191, our performance exceeds MAX-MAX. Similar conclusions are found for edge histogram and homogeneous texture.

6. Conclusion

In this paper, we explore statistical strategies of text retrieval for MIR. A term-like representation called *feature term* is proposed for media document representation, which results in an efficient query generation from multiple examples as well as an effective method of collection modelling. We adapt two text retrieval models, KL and BM25, for MIR and carry on experiments on the Corel photo collection and the TRECVid 2006 collection. This new approach brings the following benefits: (1) we are able to exploit powerful text retrieval models in multimedia domain; (2) some efficient access structures are allowed, *e.g.*inverted index, for media data processing; (3) we avoid parameter tuning in media combination and feature selection by using ranking function and aggregated features representation. Moreover, experimental results show the effectiveness of this approach, comparing with other popular methods employed in low-level feature based MIR.

7. Acknowledgement

The research leading to this paper was supported by European Commission under contracts FP6-045032 (SEMEDIA).

References

[1] Cees G.M. Snoek, Marcel Worring, Jan C. van Gemert, Jan-Mark Geusebroek, and Arnold W.M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *ACM Multimedia 2006*, 2006.

[2] Meng Yang, Barbara M. Wildemuth, and Gary Marchionini, "The relative effectiveness of concept-based versus content-based video retrieval," in *ACM MULTIMEDIA 2004*, 2004.

[3] Rong Yan and Alexander G. Hauprmann, "Query expansion using probabilistic local feedback with application to multimedia retrieval," in *CIKM 2007*, 2007, pp. 361–370.

[4] Jinxi Xu and W. Bruce Croft, "Improving the effectiveness of information retrieval with local context analysis," *ACM Trans. Inf. Syst.*, vol. 18, no. 1, pp. 79–112, 2000.

[5] ChengXiang Zhai and John Lafferty, "A risk minimization framework for information retrieval," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 31–55, 2006.

[6] Gianni Amati and Cornelis Joost Van Rijsbergen, "Probabilistic models of information retrieval based on measuring the divergence from randomness," *ACM Transactions on Information Systems (TOIS)*, vol. 20, no. 4, pp. 357–389, 2002.

[7] S.P. Harter, "A probabilistic approach to automatic keyword indexing, part i on the distribution of speciality words in a technical literature," *Journal of the ASIS*, vol. 26, pp. 197–216, 1975.

[8] E.L. Margulis, "N-poisson document modelling," in *the 15th ACM SIGIR*. ACM, 1992, pp. 177–189, ACM Press.

[9] Joo-Hwee Lim and Jesse S. Jin, "Home photo indexing using learned visual keywords," in *VIP '02: Selected papers from the 2002 Pan-Sydney workshop on Visualisation*, Darlinghurst, Australia, Australia, 2002, pp. 69–74, Australian Computer Society, Inc.

[10] Cees G.M. Snoek, Jan C. van Gemert, Theo Gevers, Bouke Huurnink, Dennis C. Koelma, Michiel van Liempt, Ork de Rooij, Koen E.A. van de Sande, Frank J. Seinstra, Smeulders, Andrew H.C. Thean, Cor J. Veenman, and Marcel WorringArnold W.M, "The mediamill trecvid 2006 semantic video search engine," in *Proceedings of the 4th TRECVID Workshop*, Gaithersburg,USA, November 2006, NIST.

[11] Apostol Natsev, Jelena Tešić, Lexing Xie, Rong Yan, and John R. Smith, "Ibm multimedia search and retrieval system," in *CIVR 2007*, New York, NY, USA, 2007, pp. 645–645, ACM.

[12] Steven C.H. Hoi, Lawson L.S. Wong, and Albert Lyu, "Chinese university of hongkong at trecvid 2006: Shot boundary detection and video search," in *TRECVid 2006 Workshop*, Maryland, USA, October 2006, NIST, pp. 76–86, NIST.

[13] TRECVID, "Analysis and presentation of soccer highlights from digital video," 2003.