

Query Generation From Multiple Media Examples

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

Abstract

This paper tries to solve the problem of query generation from multiple media examples, e.g. images, by exploiting a media document representation called feature terms. A feature term denotes a continuous interval of a media feature. This approach (1) helps feature accumulation from multiple examples to generate an efficient query; (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 and content-based 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. The 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, which creates [3]. Note that media features are of difference similarity measurements, *e.g.* euclidian distance for RGB color and intersection distance for edge histogram. This indicates a high computational complexity in query generation and document relevance estimation.

The literature related to this problem can be roughly categorised into two groups, early fusion and late fusion.

Early fusion approaches directly learn a query from extracted features. Various machine learning strategies have been proposed to decide on an optimal solution, such as active learning [2], automatic labelling [3], and super-kernel fusion [4]. However, a query example set is usually too small to support a robust analysis [5]. This seriously degrades retrieval performance if some relevant features are ignored and if some non-relevant features are over-weighted. We therefore propose an aggregation feature representation which accumulates feature from all examples. In late fusion approaches, query examples are submitted individually like a group of sub-queries. Many empirical ranking schemes are proposed to merge sub-query results to create the final output [3]. However, such a combination leads to an intensive tuning on model parameters with respect to a query [1]. As a result, it is difficult to extend late fusion approaches for general large scale MIR.

In our view, query generation is a problem of pattern mining not only from a query example set, but also in the context of a document collection. Zhai *et al.* [6] assert that the retrieval is a statistical decision problem based on the variance of term distributions in both document collection and a query. Normalising feature distribution can enlarge the significance of a query to a collection. The employment of collection knowledge alleviates the uncertainty caused by the small query example set. We develop three-step query generation from multiple examples: (1) project media features into a set of discrete variants or a vocabulary called *feature terms*; (2) define a statistic on the distribution of *feature terms* to describe the collection, as well as to create a unified query; and (3) employ text retrieval models to estimate relevance. The contributions are: (1) an efficient query generation and collection representation approach, which easily accumulates characteristics from media documents, especially low-level features; (2) a mixed ranking scheme across medias for relevance estimation, which avoids complex distance computation and parameter tuning on media/feature combination.

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

A feature term denotes a range interval of a feature. As our approach exploits the same principles employed in statistical text retrieval, we begin with a discussion about text term distribution.

As an important aspect in term weighting, text term distribution has been well discussed for the justification of retrieval models [7]. Harter *et al.* [8] propose that a term should follow a 2-Poisson distribution, because term appearance is Boolean and sparsely distributed. Margulis *et al.* [9] extend this model to N-Poisson distribution. The authors argue that N-Poisson might have provided a more precise estimation than 2-Poisson does, if a term actually followed a Poisson-like distribution. However, several class numbers from two to seven were evaluated on real document collections [9]. No specific class number of Poisson combined model shows a significant out-performance. Amati *et al.* [7] simulate a retrieval process by a Bernoulli distribution. The authors suggest a uniform term distribution, since 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 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 description¹. 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 there are lack of common concept definitions [12], and partly because of domain dependency among concepts [10]. The availability of a concept is dependent on collection domain as well as a query. Moreover, the distribution of a concept is so sparse

¹It is well known only a few dimensions of a low-level feature are effective in content representation with respect to a query [11][12].

that only a small number of documents are related with a specific concept. This results in the ineffectiveness of traditional ranking schemes. Nevertheless, the imperfection in the technique of automatic annotation reduces the reliability of concepts in retrieval [10]. Hence, low-level features are still 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 [7]. 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 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 $\hat{f} : [0, K]^N \rightarrow \{0, 1, \dots, M-1\} \sim \{0, 1\}^M$, where K denotes variable range and M the number of integers. We define these integers as *feature terms*. Since dimensions in a low-level feature are independent from each other, we take the one-dimensional case where $N = 1$.

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

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

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

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

3.1 Selection Criterion

As many methods for feature quantisation exist, some statistical criteria are necessary to decide on an optimal solution. Given the uniform assumption [7], we propose the following 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. Since the optimised term probability is $\hat{p}(t) = \frac{1}{M}$ in the uniform distribution, χ^2 test is as

follows.

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

The criterion of minimised χ^2 test is defined as,

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

Entropy measures information gain brought by a term selection.

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

A high entropy indicates a good selection.

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

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

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

$$AC = \frac{1}{M} \left\| \sum_{n=0}^{M-1} f_n e^{-\frac{2\pi i n}{M}} \right\| \quad (8)$$

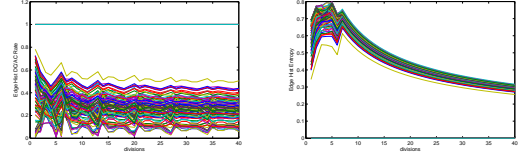
The rate of AC-DC (Equation 9) reflects the bias of the frequency sequence away from the average. A low $R_{AC/DC}$ is preferred.

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

Figure 1 a and b display criterion value distribution(y-axis) with 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



(a) ACDC Rate

(b) Entropy

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

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.

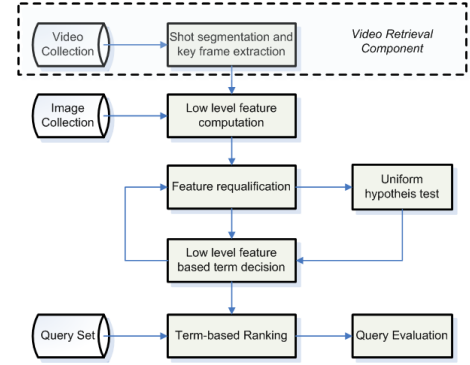


Figure 2. Retrieval System Framework

4.1 Document Representation

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

$$I_{d,V} = \{I_{t_1,d}, \dots, I_{t_n,d}\} \quad (10)$$

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 .

$$C_{Q,V} = \sum_{q \in Q} I_{q,V} \quad (11)$$

A vector representation of feature terms is defined here for document and query. Text retrieval ranking functions are used for relevance estimation.

4.2 KL Divergence Ranking

The negative KL divergence (Equation 12) [7] 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,

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

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

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

4.3 BM25 Ranking

We propose an approach based on the BM25 model for text retrieval [7]. For a media document, the frequency of a feature term t in document d , $f_{t,d}$ is the binary $I_{t,d}$. In MIR, we rely on images or keyframes from video shots. 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

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

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).

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

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

Two media collections, the Corel photo collection and the TRECVID 2006 video collection, are employed to evaluate the effectiveness of feature terms in image and video retrieval.

5.1 Corel Collection

We randomly chose 50 categories from the Corel photo collection with each category containing 100 images. Thus, a collection of 5,000 images is created for evaluation. 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, both the ground truth set and the retrieval results held 100 images. The precision and recall were therefore 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, but the BM25 ranking seems more robust.

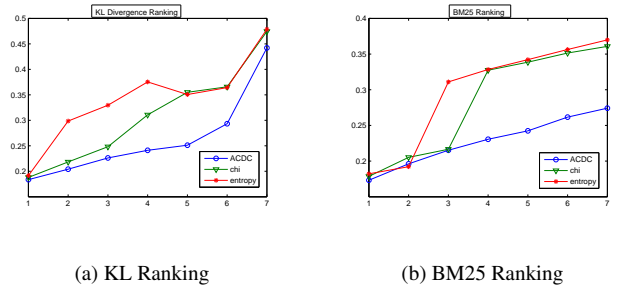


Figure 3. Retrieval Performance of EdgeHist

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.

Feature	Ranking Function	Precision at Different Example Set Size						
		1	2	3	4	5	6	7
Dominant Colour	KL	<i>0.172</i>	<i>0.265</i>	<i>0.286</i>	<i>0.381</i>	<i>0.382</i>	<i>0.384</i>	0.387
	BM25	0.036	0.138	0.136	0.135	0.135	0.150	0.150
Edge Histogram	KL	<i>0.192</i>	<i>0.299</i>	<i>0.330</i>	<i>0.375</i>	<i>0.351</i>	<i>0.364</i>	0.479
	BM25	0.182	0.192	0.311	0.328	0.342	0.356	0.370
Homogeneous Texture	KL	<i>0.179</i>	<i>0.201</i>	<i>0.219</i>	<i>0.230</i>	<i>0.242</i>	<i>0.234</i>	0.260
	BM25	0.167	0.192	0.207	0.218	0.226	0.233	0.247
Colour Layout	KL	<i>0.203</i>	<i>0.330</i>	<i>0.369</i>	<i>0.392</i>	<i>0.413</i>	<i>0.432</i>	0.548
	BM25	0.142	0.142	0.238	0.234	0.232	0.233	0.231
Colour Layout & Edge Histogram	KL	<i>0.240</i>	<i>0.332</i>	<i>0.451</i>	<i>0.488</i>	<i>0.681</i>	<i>0.722</i>	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

	Direct		kNN		Entropy		χ^2		ACDC	
	mean	Δ^2	mean	Δ^2	mean	Δ^2	mean	Δ^2	mean	Δ^2
homogenous texture	27.5	27.1	13.9	16.1	15.2	24.0	23.1	22.9	40.5	16.9
color layout	59.6	111.4	42.9	103.5	23.3	25.3	21.8	14.4	22.7	15.1
edge histogram	51.8	63.7	28.7	78.9	37.1	69.7	49.3	50.1	68.9	58.3

Table 3. Average Num-Rel-Ret of all query topics

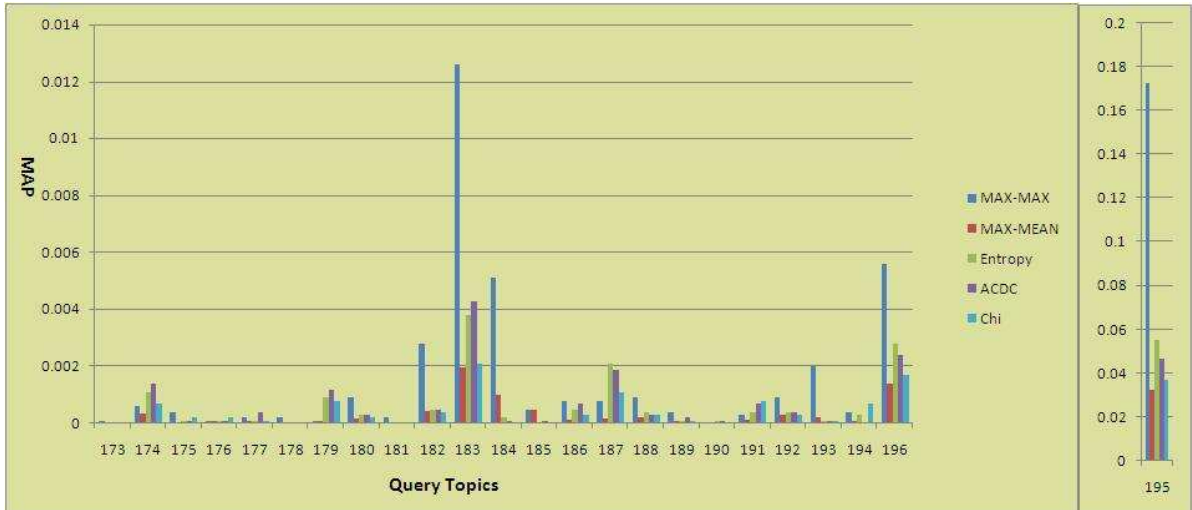


Figure 4. Mean Average Precision of Colour Layout Query

5.2 TRECVID Collection

TRECVID 2006 collection provides 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 [11]. Natsev *et al.* [12] 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.

Table 3 concludes average (mean) and standard deviation (Δ^2) of num-rel-ret for other three features. The employment of feature terms reduces num-rel-ret derivation and improves system robustness.

We also compare mean-average-precision (MAP). Figure 4 shows the MAP of colour layout in all topics. Max-Max denotes the MAP of an oracle that collects all relevant documents found by individual examples and direct comparison [11]. Max-Mean is the average MAP achieved by individual examples. Max-Max is the performance upper boundary of late fusion approaches and Max-Mean the baseline. The difference between Max-Max and Max-Mean reflects the number of examples which contribute positive knowledge. In most topics, the MAP 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. More-

over, 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 (Semia).

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. Hauptmann, "Query expansion using probabilistic local feedback with application to multimedia retrieval," in *CIKM 2007*, 2007, pp. 361–370.
- [4] Y. Wu, E.Y. Chang, K.C-C Chang, and J.R. Smith, "Optimal multimodal fusion for multimedia data analysis," in *ACM Multimedia 2004*. ACM, 2004, pp. 572–579.
- [5] 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.
- [6] ChengXiang Zhai and John Lafferty, "A risk minimization framework for information retrieval," *Inf. Process. Manage.*, vol. 42, no. 1, pp. 31–55, 2006.
- [7] 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.
- [8] 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.
- [9] E.L. Margulis, "N-poisson document modelling," in *the 15th ACM SIGIR*. ACM, 1992, pp. 177–189, ACM Press.
- [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] 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.
- [12] 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.
- [13] TRECVID, "Analysis and presentation of soccer highlights from digital video," 2003.