A BOVW Based Query Generative Model

Reede Ren[1], John Collomosse[1], and Joemon Jose[2]

[1] CVSSP, University of Surrey, Guildford, GU2 7XH, UK
[2] IR Group, University of Glasgow, Glasgow, G12 8QQ, UK

Abstract. Bag-of-visual words (BOVW) is a local feature based framework for content-based image and video retrieval. Its performance relies on the discriminative power of visual vocabulary, *i.e.* the cluster set on local features. However, the optimisation of visual vocabulary is of a high complexity in a large collection. This paper aims to relax such a dependence by adapting the query generative model to BOVW based retrieval. Local features are directly projected onto latent content topics to create effective visual queries; visual word distributions are learnt around local features to estimate the contribution of a visual word to a query topic; the relevance is justified by considering concept distributions on visual words as well as on local features. Massive experiments are carried out the TRECVid 2009 collection. The notable improvement on retrieval performance shows that this probabilistic framework alleviates the problem of visual ambiguity and is able to afford visual vocabulary with relatively low discriminative power.

1 Introduction

Bag-of-visual words (BOVW) is a promising framework for content-based image and video retrieval. Key points are collected from images to denote salient regions; local features such as SIFT [8] are computed around key points and are clustered to generate a visual vocabulary; each SIFT cluster is registered as a unique visual word; an image is therefore translated into a set of visual words; appearance statistics, *e.g.* the histogram of visual word frequency [5], is calculated to decide on the contribution of visual words as well as to estimate the relevance. Fei-Fei *et al.* [6] value BOVW as a reliable approach to bridge the gap between low-level visual characters and high-level image contents. They demonstrate the effectiveness in general scene discrimination and visual object modelling. The recent TRECVid competition also reports that BOVW achieves the best performance in the automatic search task, although the overall precision remains unsatisfied [12,13].

One of the major challenges in BOVW is the generation of visual vocabulary. Since visual words do not naturally exist in images, vocabulary generation is an optimisation across the local feature set of the document collection, which tries to maximise the discriminative power while not reduces retrieval efficiency. Given the extremely huge number of local features and the sparsely distributed visual contents, such an optimisation process is of a high complexity and is not so robust

K.-T. Lee et al. (Eds.): MMM 2011, Part I, LNCS 6523, pp. 118–128, 2011.

[13]. K-mean clustering and its variants are widely adopted as a low cost solution but requires an extra preliminary of vocabulary size, *i.e.* cluster number [16]. Jiang *et al.* [5] carried out massive experiments on the TRECVid 2006 collection and evaluated various vocabulary size from 200 to 10,000. The authors observe the problem of visual ambiguity that a local feature can be assigned to multiple visual words and report no vocabulary configuration out-performs the others. This indicates that visual ambiguity weakens the effectiveness of visual words for content discrimination and that the visual vocabulary generated by K-mean clustering is unreliable for retrieval [12]. Moreover, appearance statistics is closely associated with the vocabulary. In addition, Li *et al.* [6] point out that a high-level visual concept is contributed by a set of rather than a single local feature. An improperly configured vocabulary not only misguides appearance statistics but also leads to the problem of 'double counting' in pattern recognition, as a visual concept is randomly projected onto multiple visual words.

This paper aims to relax the constrains from visual vocabulary and thus to alleviate related problems such as visual ambiguity and visual word assignment. We propose the probabilistic framework of query generative model for BOVW based retrieval. Latent query topics are learnt directly from local features. This avoids the usage of visual vocabulary in query generation and results in a precise query modelling. Moreover, local feature based query models allows the estimation of visual word contributions to a local feature. A new soft assignment scheme is therefore developed by exploiting neighbourhood statistics of visual words around a local feature. Experimental results show that this new assignment scheme is more effective than cluster centre based assignment. In addition, the query generative model make possible the introduction of prior/external knowledge, *i.e.* local feature based visual concept model, and promises a further improvement on retrieval performance. Some research issues are also addressed, *e.g.* shot-based temporal accumulation.

The remainder of this paper is organised as follows. Section 2 briefly reviews components in the framework of BOVW, including key point detection, visual vocabulary generation and relevance estimation. Section 3 explains the query generative model and describes the estimation of (1) local feature distribution, (2) latent topic distribution among query examples; and (3) local feature distribution in a latent topic. The optimised number of query topics is learnt by maximising the entropy of latent topic distribution. Experimental results are stated in Section 4, including experimental configuration, baseline creation and shot-based retrieval performance on the TRECVid 2009 collection. Conclusions are found in Section 5.

2 Related Work

In this section, we briefly review the development of BOVW in content-based video retrieval. BOVW was originally proposed for visual object recognition [4]. Snoek *et al.* [12] introduce this framework into content-based video retrieval and treat detection confidence as an effective relevance estimator. The authors claim that BOVW is a possible gateway to the robust semantic video retrieval. Three key issues are then identified in [5], including key point detection, vocabulary generation and visual word weighting scheme.

Key points are essential small image regions for the recognition of visual contents. Detection algorithms such as Harris Laplace, Boost Colour Harris Laplace, Difference of Gaussian (DOG) and grid sampling are compared in [13]. Uijlings *et al.* [13] justify their choice by retrieval performance on the TRECVid 2006 collection. The authors recommend grid sampling for discriminative power despite the largest key point collection among all. We notice that the performance difference is insignificant to support such a computational cost. Therefore, we use the DoG key point detector [8] in this paper to balance the effectiveness and the efficiency. Later, Battiato *et al.* [2] group key points to form visual synset, the counterpart of n-gram representation in textual documents. The authors show that visual synset is more invariant and more discriminative than a single key point in object categorisation. However, the collection of possible visual synsets increases exponentially with the average synset length. Liu *et al.* [7] explore Ababoost to select the most discriminative combinations. Savarese *et al.* [11] compute correlation matrixes to identify the most common key point co-occurrence. Zhang *et al.* [16] apply geometric constrains to group key points nearby. These works decreases the blooming speed to sub-exponent.

Vocabulary generation is usually regarded as a problem of unsupervised learning, which is closely associated with two factors: (1) the distribution of visual concepts and (2) the distribution of local features (Figure 1b). The number of visual concepts decides on the necessary size of a visual vocabulary, although this requirement is latent. This is due to the complexity in the estimation of the actual concept number in general purpose retrieval. Local feature distribution defines the boundary between visual words and therefore affects the discriminative power of vocabulary. Marsezlek *et al.* [9] propose K-mean clustering with a random cluster number for efficiency. Uijling *et al.* [13] adapt the random forest as a robust replacement for K-mean. Some dimensionality removal approaches are also tried to improve the effectiveness, *e.g.* principle component analysis [13]. However, it remains a research question how to optimise a visual vocabulary, given the huge feature collection.

BOVW adapts textual term weighting schemes for relevance estimation, *e.g.* term frequency and inverse document frequency. Evering *et al.* develop a binary weighting for image based visual word matching. Marsezlek et al. [9] adapt a BM25 like retrieval model to estimate the relevance between video shots. Agarwal *et al.* [1] notice visual ambiguity and take the difference between local features into consideration. The authors propose a language model like probabilistic framework to simulate local feature distribution as well as to optimise the vocabulary by using the posterior mixture-component membership. Jiang *et al.* [5] propose soft assignment and label a local feature with multiple visual words at different weights. This results in a significant precision improvement on the TRECVid 2006 collection. Germert *et al.* [14] furthermore propose a query dependent soft assignment in which the weight of a local feature to a visual word is learnt from query classes.

3 Query Generative Model

In this section, we justify our query generative framework for BOVW based retrieval. BOVW can be regarded as a four-tier document generative model (Figure 1b). A visual word denotes a distribution of local features; a visual concept is represented by a probabilistic distribution across visual words [6]; the relevance is estimated by comparing the mixed distribution of visual concepts [12]. In addition, this process can be simplified into three layers by projecting local features directly onto visual concepts, *e.g.* the visual concept ontology called ImageNet[3]. As a consequence, the relevance estimation in BOVW is to generate a visual document by picking a distribution over visual concepts and then a concept dependent distribution of visual words/local features (Figure 1a).

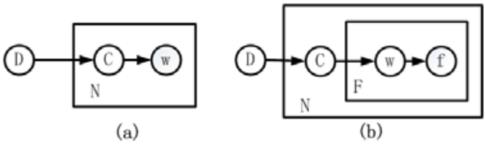

Fig. 1. BOVW based generative model, where **D** denotes visual documents, **C** stands for visual concepts, **w** and **f** are visual words and local features, respectively

Document generative models such as Latent Dirichlet Allocation (LDA) are of high complexity. Given the huge number of local features, it is computationally unfeasible to exploit document generative models. We hence turn to an alternative called query generative model (Figure 2). Zhai *et al.* [15] prove that the document generative model is theoretically equivalent to the query generative approach. Moreover, there are three advantages in the query generative model. Firstly, effective query concepts are densely distributed in query example collection. This facilitates the learning of latent visual concepts. Secondly, query example collection is small. It is unnecessary to define an individual vocabulary to conceal local feature variations. This alleviates the problem of vocabulary generation and indicates that the query generative model can afford visual vocabulary of relatively low discriminative power. Finally, the query generative model provides a precise modelling for relevance and allows the usage of prior knowledge, such as local feature based concept models in ImageNet. This will improve the effectiveness and the robustness of BOVW based retrieval. In summary, the query generative model provides a dynamic probability framework for BOVW based relevance estimation.

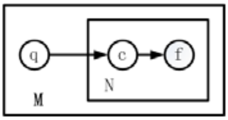

Fig. 2. Query generation model, where q denotes a query example, C and f are visual concepts and local features, respectively

The relevance estimation function (RSV) is as follows.

$$\begin{aligned} p(d|Q) &= \sum_{C_Q} p(d|C_Q)p(C_Q|Q) \\ &= \sum_{C_Q}(\sum_{f_Q} p(d|f_Q)p(f_Q|C_Q))p(C_Q|Q) \end{aligned}$$

where C_Q denotes latent concepts in a query Q; d stands for a video document; f_Q and f_d are local feature collection of Q and d, respectively. This function shows that checking the most relevant query concepts is an effective alternative to studying all possible concepts in documents. If local features are quantified into visual words, the RSV can be rewritten into the joint set description.

$$p(d|Q) = \sum_{C_Q}(\sum_{f=f_Q \cap f_d} p(d|f)p(f|C_Q))p(C_Q|Q) \tag{1}$$

where f_d is the local feature collection in a visual document. This description is equivalent to the binary/tf-idf based weighting scheme, where $p(d|f_Q)$ stands for the binary matching between visual words. Consider visual ambiguity in the BOVW framework, f_d can be treated as a noisy derivation from f_Q. We hence improve the RSV as follows.

$$p(d|Q) = \sum_{C_Q}(\sum_{f_Q}(\sum_{f_d} p(d|f_d)p(f_d|f_Q))p(f_Q|C_Q))p(C_Q|Q) \tag{2}$$

$p(f_d|f_Q)$ is a similarity measurement, which can be an assignment scheme to highlight the contribution of local feature f_Q to visual word f_d. For example, the soft-assignment in [5] is an Euclidean distance based contribution estimation, where $p(f_d|f_Q) = \frac{1}{2^{L_2(f_d, f_Q)}}$.

To summarise, three components are to learn in the query generative approach, the concept distribution in query examples $p(C_Q|Q)$, the local feature distribution for a concept $p(f_Q|C_Q)$ and the local feature distribution in documents $p(d|f_d)$. This is because $p(d|f_d) = \frac{p(f_d|d)p(d)}{p(f_d)} \sim p(f_d|d)$, as d is uniformly distributed and $p(f_d)$ is constant. In addition, the RSV (Equation 2) indicates that the query generative approach will be more effective in shot-based video retrieval. This is because $p(f_d|d)$ would be a Boolean if only one keyframe were considered.

3.1 $p(f_d|f_Q)$

$p(f_d|f_Q)$ is a similarity measurement between local features in visual documents and in query examples. The approximation of cosine distance is ineffective because local feature descriptors, such as SIFT [8], are a combination of edge and texture histogram. Note that p(fd—fQ) could be regarded as a soft assignment scheme. We consider the ad-hoc density of local feature distribution (Figure 3). This alleviates the preliminary requirement on vocabulary size.

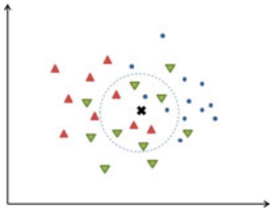

Fig. 3. Neighbourhood based soft assignment, where x denotes f_Q

We project f_Q directly into the local feature space and collect a given size of neighbourhood, e.g. 0.001% of local feature space. The appearance number of a visual word in the neighbourhood is counted and the appearance ratio is taken as the soft assignment weight.

$$p(f_d|f_Q) = \frac{\#_R(f_d)}{\#_{f_d \in R}(f_d)} \tag{3}$$

where R is the neighbourhood of f_Q. The usage of pyramid index ensure this computation of a constant complexity.

3.2 $p(f_d|d)$

$p(f_d|d)$ counts the appearances of local features in a video document d. Since not all SIFTs could be matched frame by frame, we develop a two-pass strategy to improve the robustness, including SIFT matching and block-tracking. SIFT matching compares 128-dimensional SIFT values to couple local features. If not matched, we turn to the step of block tracking. A visual frame is re-sized by affine parameters in SIFT descriptor. A $4 * 4$ block is collected around the point and searched in nearby areas of sequent frames. If the colour-based block distance is small enough, we think this local feature remains though ignored by the key point detector. Hence, $tf = tf_{sift} + atf_{tracking}$, where tf_{sift} and $tf_{tracking}$ are matching count by sift matching and block tracking, respectively; a is 0.75 learnt by experiments. $p(f_d|d)$ is estimated as follows.

$$p(f_d|d) = \frac{tf}{\|d\| + s} + \epsilon \tag{4}$$

where $\|d\|$ is the number of key points in a shot, s denotes the size of low-level feature terms [10] and ϵ is a smoothing parameter. Two types of feature terms, colour layout and edge histogram, are used in experiments to describe global characters.

3.3 $p(C_Q|Q)$ and $p(f_Q|C_Q)$

$p(C_Q|Q)$ and $p(f_Q|C_Q)$ aim to fit a given number of visual concepts to query examples. In the TRECVid collection, a query consists of 7-11 images and a short textual description. We take the number of textual noun entities as the initial estimation of possible concept number. The example set is re-sampled to create a serial of constant size sub-query collections $(q_1 \dots q_M)$. We compute $p(f|q_i)$ and Bootstrap is used to improve the estimation. $p(f|q_i)$ is computed by the EM algorithm.

E-Step

$$p(c_l|q_i, f_j) = \frac{p(f_j|c_l)p(c_l|q_i)}{\sum_{l=1}^{K} p(f_j|c_l)p(c_l|q_i)} \tag{5}$$

M-Step

$$p(f_j|c_l) = \frac{\sum_{i=1}^{M} tf_j(q_i)p(c_l|q_i, f_j)}{\sum_{j=1}^{N}\sum_{i=1}^{M} tf_j(q_i)p(c_k|q_i, f_j)} \tag{6}$$

$$p(c_l|q_i) = \frac{\sum_{j=1}^{N} tf_j(q_i)p(c_k|q_i, f_j)}{\sum_{j=1}^{N} tf_j(q_i)} \tag{7}$$

where N is the number of local features and K the number of relative concepts. The sum of feature entropy is calculated to decide on the number of effective concepts. In addition, the actual query concept number is less than five in the TRECVid 2009 collection.

$$\|C_Q\| = \arg\max_{K} \sum_{l=1}^{K}\sum_{j=1}^{N}(-p(f_j|c_l)\log p(f_j|c_l)) \tag{8}$$

4 Experiment

The TRECVid 2009 collection is used for evaluation, which consists of 219 videos and 21 queries. Every video is made up by about 150 shots and the overall shot number is 212,256. We sample a shot at 1/10. DOG is used to detect key points and 128-dimensional SIFT to describe local characters. The SIFT collection contains about 17M samples. Retrieval performance is measured by the TRECVid evaluation tool [12]. The following sections will address the configuration of retrieval system, including soft assignment scheme, shot-based relevance and visual vocabulary generation, and finally compare the performance with LDA and the best record in the competition.

4.1 Soft Assignment

Soft assignment is an efficient approach to alleviate visual ambiguity. In this paper, we exploit the neighbourhood statics around a local feature. There are two issues affecting the effectiveness: neighbourhood scope and vocabulary size. The choice of neighbourhood is a trade-off between the efficiency and statistical robustness. On one hand, a large neighbourhood ensures statistical robustness at the cost of efficiency by reaching more samples than a small one does. On the other hand, a small neighbourhood needs less efforts in indexing but requires an extra smoothing to avoid too many zero weights.

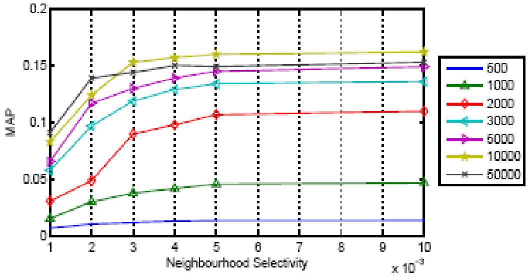

Fig. 4. MAP over neighbourhood selectivity and vocabulary size

Figure 4 compares the retrieval performance with numerous neighbourhood scope from 0.0001% to 0.01% of the SIFT collection space under all vocabulary configurations. Experimental results show that a large neighbour is able to increase precision, although the improvement will diminish after the neighbourhood is larger than 0.003%. This shows the neighbourhood based soft assignment is robust and effective. Another interesting observation is about vocabulary size. Figure 4 shows that a large vocabulary also improves retrieval performance even though a small neighbourhood is used. This consistence indicates that neighbour based assignment can be improved by a discriminative vocabulary and that the query generative model alleviates the problem of visual ambiguity. In addition, it costs about 17 milliseconds to collect samples of 0.005% feature space in pyramid index. We therefore use the neighbourhood of 0.005% to compute the soft assignment weight and the vocabulary of 10,000 in the following experiments.

Table 1 compares two soft assignment schemes, distance-based [5] and neighbourhood based. Neighbourhood based scheme is significantly better than distance-based.

4.2 Shot-Based Relevance

An obvious difference between video and image retrieval is temporal continuity inside media documents. Many video retrieval systems regard a shot as an image collection and represent a shot by accumulating local features from key frames. Figure 5 compares performances under various weights from 0.4 to 1 on local

Table 1. MAP over soft assignment scheme

Visual word number	Distance-based @ 5 classes [5]	Neighbourhood-based @ 0.005%
500	0.008	0.014
1000	0.010	0.046
2000	0.014	0.107
3000	0.015	0.134
5000	0.019	0.145
10,000	0.021	0.167
50,000	0.012	0.149

feature tracking. A 4.38% improvement is observed over the best performance on key frame based retrieval (10,000 visual words and neighbourhood of 0.005% feature space). However, an improper weight may also decrease the performance. This indicates that the improvement from temporal tracking might be limited. There are two possible explanations: (1) an intensive key frame sampling is good enough for shot representation in retrieval and (2) query topics in the TRECVid 2009 mostly focus on a static scene rather than a continuous motion.

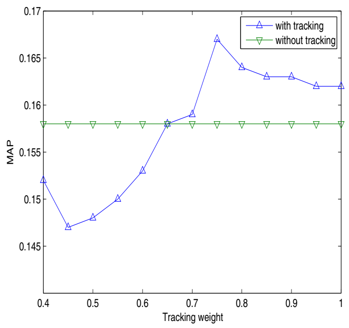

Fig. 5. MAP over shot accumulation

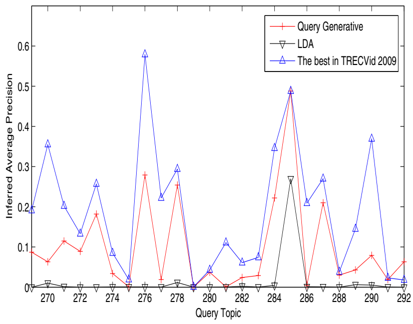

Fig. 6. Top AP in the TRECVid 2009 collection

4.3 Retrieval Performance

We use the latent Dirichlet allocation (LDA) as the baseline [10]. LDA is one of the most widely used generative model and finds applications in textual segmentation, multi-class face classification and spoken letter recognition. We use the TRECVid 2008 collection to train LDA with the latent topic number 50. 10,000 visual words are assigned to each latent topic by using reserve-jump Markov Monte Carlo. Then, a query is projected onto latent topics and the relevance is estimated by maximising the post-probability that a document to the latent topic distribution. The Merry run in the MediaMill [12] is also used as baseline due to the similarity in retrieval configuration, *e.g.* BOVW representation, feature-based retrieval and learning on-the-fly query concepts. The difference between Merry run and ours is the estimation of relevance and the scheme of soft assignment. The Merry run casts query topics onto a group of query classes and uses support vector machine to rank documents. Table 2 compares retrieval performance by LDA, the query generative model, the Merry run in MediaMill [12] and the best-over-all in the TRECVid 2009 [10]. The best-over-all collects the best performance on every topics in the TRECVid 2009 competition, which exploits all modalities including audio scripts, high-level concepts and textual tags.

Table 2. TRECVid 2009 Performance

RUN	MAP
LDA [10]	0.0132
Query Generation	0.167
Merry Run [12]	0.089
Best over all [10]	0.188

Topic based average precision is shown in Figure 6.We achieve high scores on scene related topics, *e.g.* people at a table with a computer visible and building entrance. This shows the effectiveness of the query generative scheme in the description of multiple visual concepts.

5 Conclusion

In this paper, we adapt the query generative model for BOVW based video retrieval. This robust probabilistic framework incorporates latent visual concepts and exploits visual word distribution on local features for a superior retrieval performance. Its computation cost is lower than document generative models. This makes possible the exploitation of BOVW based generative models in a large scale retrieval. Moreover, the query generative model projects local features directly onto latent visual concepts. This alleviates the problem of visual ambiguity and relaxes the requirement on the discriminative power of visual vocabulary. System robustness is therefore improved. Nevertheless, this new BOVW-based retrieval framework facilitates the introduction of prior visual concept models into relevance estimation. This extends the scope of BOVW and allows the usage of web-based external knowledge. In short, a further improvement will be seen on retrieval performance.

References

1. Agarwal, A., Triggs, B.: Multilevel image coding with hyperfeatures. International Journal of Computer Vision 78(1), 15–27 (2008)
2. Battiato, S., Farinella, G.M., Gallo, G., Ravì, D.: Spatial hierarchy of textons distributions for scene classification. In: Huet, B., Smeaton, A., Mayer-Patel, K., Avrithis, Y. (eds.) MMM 2009. LNCS, vol. 5371, pp. 333–343. Springer, Heidelberg (2009)

3. Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR 2009 (2009)
4. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106(1), 59–70 (2007)
5. Jiang, Y.-G., Ngo, C.-W.: Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. Computer Vision and Image Understanding 113(3), 405–414 (2009)
6. Li, L.-J., Socher, R., Fei-Fei, L.: Towards total scene understanding:classification, annotation and segmentation in an automatic framework. In: Proc. IEEE Computer Vision and Pattern Recognition, CVPR (2009)
7. Liu, D., Hua, G., Viola, P., Chen, T.: Integrated feature selection and higher order spatial feature extraction for object categorisation. In: CVPR 2008, pp. 1–8 (2008)
8. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV, pp. 1150–1157 (September 1999)
9. Marszalek, M., Schmid, C., Harzallah, H., van de Weijer, J.: Learning representations for visual object class recognition. In: ICCV (2007)
10. Punitha, P., Misra, H., Ren, R., Hannah, D., Goyal, A., Villa, R., Jose, J.M.: Glasgow university at trecvid 2009. In: TRECVID (2009)
11. Savarese, S., Winn, J., Criminisi, A.: Discriminative object class models of appearance and shape by correlatons. In: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2006, Washington, DC, USA, pp. 2033–2040. IEEE Computer Society Press, Los Alamitos (2006)
12. Snoek, C.G.M., van de Sande, K.E.A., de Rooij, O., Huurnink, B., van Gemert, J., Uijlings, J.R.R., He, J., Li, X., Everts, I., Nedovic, V., van Liempt, M., van Balen, R., de Rijke, M., Geusebroek, J.-M., Gevers, T., Worring, M., Smeulders, A.W.M., Koelma, D., Yan, F., Tahir, M.A., Mikolajczyk, K., Kittler, J.: The mediamill TRECVID 2009 semantic video search engine. In: TRECVID (2009)
13. Uijlings, J.R.R., Smeulders, A.W.M., Scha, R.J.H.: Real-time bag of words, approximately. In: CIVR 2009, Santorini, Fira, Greece, pp. 1–8. ACM, New York (2009)
14. van Gemert, J.C., Veenman, C.J., Smeulders, A.W., Geusebroek, J.-M.: Visual word ambiguity. IEEE Transactions on Pattern Analysis and Machine Intelligence 32, 1271–1283 (2010)
15. Zhai, C., Lafferty, J.: A risk minimization framework for information retrieval. Inf. Process. Manage 42(1), 31–55 (2006)
16. Zhang, S., Tan, Q., Hua, G., Huang, Q., Li, S.: Descriptive visual words and visual phrases for image applications. In: ACM Multimedia 2009 (2009)