Entropy based Supervised Merging for Visual Categorization

Usman Farrokh Niaz and Bernard Merialdo

EURECOM, 2229 Route des Cretes, 06560 Sophia Antipolis, France

Abstract. Bag Of visual Words (BoW) is widely regarded as the standard representation of visual information present in the images and is broadly used for retrieval and concept detection in videos. The generation of visual vocabulary in the BoW framework generally includes a quantization step to cluster the image features into a limited number of visual words. This quantization achieved through unsupervised clustering does not take any advantage of the relationship between the features coming from images belonging to similar concept(s), thus enlarging the semantic gap. We present a new dictionary construction technique to improve the BoW representation by increasing its discriminative power. Our solution is based on a two step quantization: we start with k-means clustering followed by a bottom-up supervised clustering using features' label information. Results on the TRECVID 2007 data [8] show improvements with the proposed construction of the BoW.
We equally give upperbounds of improvement over the baseline for the retrieval rate of each concept using the best supervised merging criteria.

1 Introduction

The codebook or Bag of Words (BoW) model is a histogram representation used for scene description that is proven to be promising for large scale image and video retrieval. It is usually obtained through vector quantization performed on a number of keypoints or robust descriptors gathered from images. Each image is in turn coded by this histogram representation.

For the generation of visual vocabulary in the BoW framework, keypoints or Local Interest Points (LIPs) containing rich local information from images are identified. These keypoints are described using local image descriptors such as Scale Invariant Feature Transformation (SIFT) [5] resulting in a 128 dimensional feature vector and are clustered to form the visual codebook. This quantization is usually performed using any unsupervised clustering algorithm, like e.g. k-means. The clustering process divides the feature space into adjacent Voronoi cells where the cluster centers are the words of the visual vocabulary. After quantization of the feature space an image can be represented by a histogram where the bins of this histogram count the number of visual words in each cell. This histogram is then used for training the classifier; typically the two class Support Vector machines (SVM).

Image description in the BoW framework generally faces two important issues. The first is that generating the visual vocabulary through unsupervised quantization from tens of thousands of low level descriptors does not capture semantic context as category information is not accounted for during clustering. Doing so the expressive or discriminative power of the vocabulary is affected as only overall distortion is minimized and category information is not used increasing the semantic gap between the concept and the mid-level BoW feature. This category information should be used in the vocabulary generation to build class specific visual words. The other problem with codebook representation is choosing the vocabulary size. Typical size of visual vocabulary ranges from 200 to 5000 words. The categorization performance usually increases with the dictionary size but this affects the retrieval efficiency and also the generalization ability of the vocabulary over noisy descriptors. There is therefore a need to find a compromise between the dictionary size and its discrimination ability. We present a dictionary construction method in this paper, to address these issues, for generating discriminative codebooks to improve retrieval results.

The problem of increasing the discriminative power of BoW model has been attacked by many authors in the recent years. Wang [9] builds a multi-resolution codebook by adding a new codeword at each step using hierarchical clustering and a selection criterion based on Boosting. This is done to find a compromise between a small codebook that lacks discriminative power and a large one that may result in overfitting. Perronnin et al. [7] represent each image with a bipartite histogram by building universal and class specific vocabularies using maximum likelihood estimation. Lin et al. [4] use a similar principle to bridge the semantic gap between the concept(s) depicted in the image and the low level features. k-means is used to generate separate class specific vocabularies followed by an agglomerative clustering on class codebooks to get the uinversal vocabulary. In both these works an image is represented by a set of histograms, one per class, using the amalgamated codebooks. Hao and Jie [3] present an improved BoW algorithm for scene recognition exploring discriminative power of codewords when representing different scene categories. They obtain a weighted histogram to code every image that highlights the discriminative capabilities of each codeword for each category.

For generating a discriminative codebook we follow a two step clustering framework as proposed by Winn et al. [10], where they compress an initial large dictionary by optimizing a statistical measure of discrimination that finds a compromise between low intra-class variance and inter-class discrimination. Moosmann et al. [6] build a set of randomized decision trees using the class labels with the leaf of a tree representing a spatial code (visual word). They calculate information gain of the split at each step of tree growing and use it as a threshold to split the tree based on the descriptor dimension at that level. Similarly, we use en entropy measure to merge clusters, achieved through an initial clustering, by minimizing information loss.

We use a clustering method with only a few k-means' mean shift iterations using a better centers initialization based on [1] to generate a larger than required

number of clusters before doing a supervised mapping significantly reducing the number of clusters (visual words). We initially merge neighboring clusters based on entropy minimization criteria that allows the generation of non-convex connex clusters. We present three such merging criteria in order to increase the discriminative power of BoW with an increase in the retrieval performance over the baseline. We then relax certain constraints in our merging criteria to allow the generation of non-connex clusters. We have used SIFT descriptors [5] calculated on keypoints extracted from images, contrary to dense sampling [10, 6], labeled with one or more classes rather than segmented hand-labeled images [10].

We also show that using our dictionary construction from supervised merging a smaller dictionary gives the performance comparable to the retrieval performance given by a dictionary upto 8 times its size.

The rest of this paper is organized as follows. Section 2 gives the detailed description of the two step supervised clustering algorithm and its three variants. In Sect. 3 we discuss detailed experimentation and present the results with the improvements proposed. Finally Sect. 4 concludes the paper.

2 Supervised Clustering Based on Entropy Minimization

In the two step clustering paradigm, Fig. 1, first of all the nearest neighbors are quantized into a large number of visual words using k-means. The number of initial visual words (k-means clusters) is $p * D$, where D is the size of the desired dictionary. In the second step the number of clusters is reduced by $1/p$ by merging neighboring clusters repeatedly based on entropy driven information loss minimization criterion.

2.1 Concept Distribution Entropy Minimization

For deriving this minimization criterion we have the m concepts $X_l \in \mathbf{X}, l = 1 \dots m$ and we know with what concept(s) is each image I labeled. Thus we know what concept is represented by each descriptor (keypoint). Now suppose as a result of the initial clustering we have $p * D$ clusters, and for each cluster $C_i \in \mathbf{C}, i = 1 \dots p * D$ we know the labels of the keypoints assigned to it (we are only treating keypoints coming from labeled shots). For finding the number of keypoints belonging to a concept X_k in the cluster C_i we consider that there may exist shots that are labeled with more than one concept. Also generally keypoints extracted from a shot are assigned to different centers. Thus we compute the number of occurences of concept X_k in the cluster C_i as:

$$|X_k \in C_i| = \sum_{I labeled with X_k} \frac{|Keypoints(I) \in C_i|}{|Keypoints(I)|} \tag{1}$$

We find next the conditional probability of the concept X_k given the cluster C_i:

$$p(X_k/C_i) = \frac{|X_k \in C_i|}{\sum_l |X_l \in C_i|} \tag{2}$$

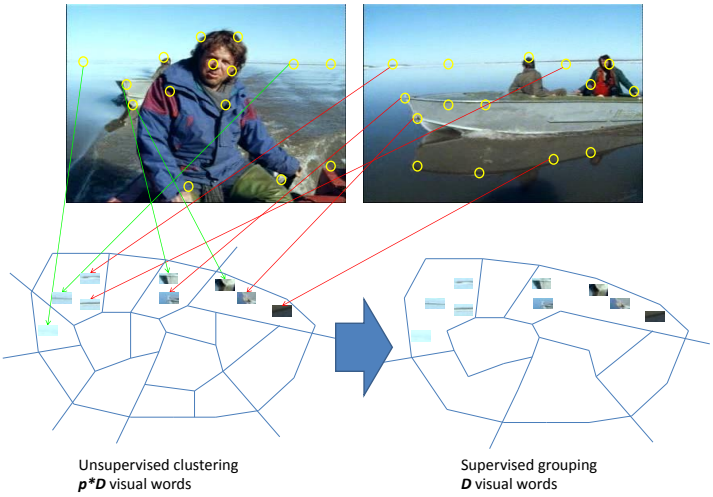


Fig. 1: Supervised merging of visual words

The set $Nb\{i\}$ contains the neighbors of the cluster C_i where two clusters are neighbors if the midpoint between their centers is closer to those two centers than to any other center. When joining two neighboring clusters C_i and C_j, where $j \in Nb\{i\}$, all the keypoints in C_j are assigned to C_i, and all the neighbors of C_j are added to those of C_i. C_j is then deleted from the set of clusters i.e. $C = C \backslash C_j$.
The entropy of the concept distribution given a clustering is given by:

$$H(X/C) = -\sum_{C} p(C) \sum_{X} p(X/C) \log p(X/C) \tag{3}$$

which is increased (or stays the same) when any two clusters are merged.

The combination $C_i \cup C_j$ that minimizes the increase in entropy is our target combination and those two clusters are merged together.

$$\underset{C_i \cup C_j, j \in Nb\{i\}}{\operatorname{argmin}} \quad H(X/C) \tag{4}$$

Thus the entropy $H(X/C)$ is calculated for a given clustering C. This step is repeated $p * D - D$ times until the desired number of clusters D is reached.

2.2 Concept Dependent Entropy Minimization

The entropy minimization principle can be equally used to find a merge of clusters independently for each concept using entropy of only that concept. This way we shall have one combination of clusters per concept and thus we will end up with a different BoW representation for each semantic concept. Using the above

notation, for the concept $X_k \in \mathbf{X}$ the entropy is given by:

$$H^{cd}(X_k/C) = -\sum_{C \in \mathbf{C}} \left[p(X_k, C) \log p(X_k/C) + p(\overline{X_k}, C) \log p(\overline{X_k}/C) \right] \quad (5)$$

where $p(\overline{X_k}, C) = p(C) - p(X_k, C)$ and $p(\overline{X_k}/C) = 1 - p(X_k/C)$.

Now for each possible combination of two neighboring clusters we will calculate the entropy to find the best merge by choosing the two clusters that result in minimum entropy increase, given by:

$$\underset{C_i \cup C_j, j \in Nb\{i\}}{\operatorname{argmin}} H^{cd}(X_k/C) \quad (6)$$

This step is repeated, reducing the total number of clusters by one each time, until the desired number of clusters is reached. This whole process is repeated for each concept resulting in a different clustering for each concept as well as a different bag of words model. An image is thus represented by a set of histograms, one per concept.

2.3 Average Concept Entropy Minimization

Another possibility to obtain a clustering combination is by combining the output of the concept dependent clusterings. This is done by taking the sum of entropy of all concepts for a merge of two clusters and then minimizing that sum for every possible combination of clusters. This clustering of average over all concepts is given by:

$$\underset{C_i \cup C_j, j \in Nb\{i\}}{\operatorname{argmin}} \sum_{X_k \in \mathbf{X}} H^{cd}(X_k/C) \quad (7)$$

where $H^{cd}(X_k/C)$ is the concept dependent entropy as given in (5).

2.4 Relaxing Constraints

Based on the results shown in Sect. 3 we select the best entropy minimization based clustering criterion and make few changes. To reduce the bias of the labeled keframes over the unlableled ones we include all the keypoints in the second step of clustering. This is done by including all the unlabeled keypoints as the $(m+1)^{th}$ concept during the calculation of the entropy of the concept distribution and recalculating the mapping based on entropy minimization. Furthermore we relax the constraint of merging only neighboring clusters where any two clusters (not necessarily neighbors) can be mapped together in the high dimensional disjoint clustering space. This allows the generation of a non-connex BoW model. These alterations are further explored in the Sect. 3 discussing the experiments and results.

Table 1: Mean Average Precision for 20 concepts using three entropy minimization based mapping criteria

Dictionary Size	***K-means***	**Min Ent**	**Av Cd**	**Cd**
500	0.0739			
1000 to 500		0.0795	0.0792	0.0757
2000 to 500		0.0801	0.0791	0.0758
4000 to 500		0.0813	0.0775	0.0727

3 Experiments

We present here experiments carried out on the TRECVID 2007 Sound and Vision database comprising 219 videos [8]. The training corpus consists of 110 videos and the other half is used for tests. Twenty semantic concepts are used to demonstrate the results. We have used 1 vs all SVM classifiers with chi-square kernel of degree 2 using the LIBSVM [2] package for each concept.

3.1 Supervised Clustering Results

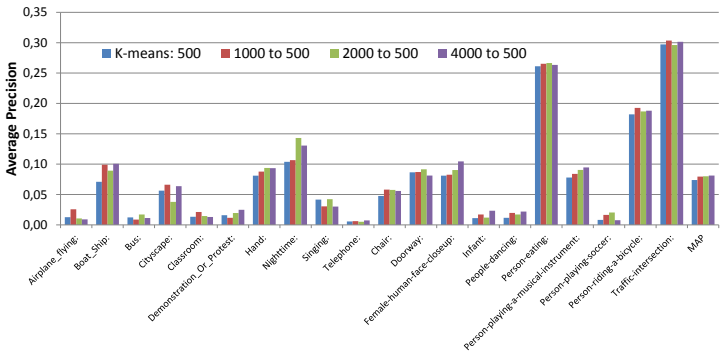
Initially we evaluate the performance of supervised clustering for a resulting dictionary of 500 visual words using the three types of entropy minimization criteria. In our experiments the maximum value of p, as described in Sect. 2, is 8. That is the maximum size of initial visual dictionary obtained through k-means is 8 times the size of the desired supervised dictionary. To obtain a large initial dictionary we have used **k-means++** algorithm [1] for a better initialization in order to avoid a large number of k-means iterations, which is costly for a large number of centers. K-means++ is an initialization method that selects initial seeds far from each other while minimizing the effects of outliers. This is done by chosing a new cluster centers with a probability propotional to its distance to the closest center already chosen. After the initialization 10 normal k-means iterations are performed to generate initial visual dictionaries.

Using these large dictionaries supervised mappings are done from clustering space with 1000, 2000 and 4000 centers to 500 centers by merging neighboring clusters using entropy minimization criteria. In all three cases the final dictionary generated is always 500-word big which is then used to represent images as histograms. SVM classifiers are trained for each concept and independently for each set of histograms obtained through entropy minimization based mappings. The Mean Average Precision (MAP) for all 20 concepts is shown in the Table 1 for the 3 mapping criteria and for the 3 initial cluster sizes, along with the MAP obtained using 500 visual words achieved directly through k-means (baseline).

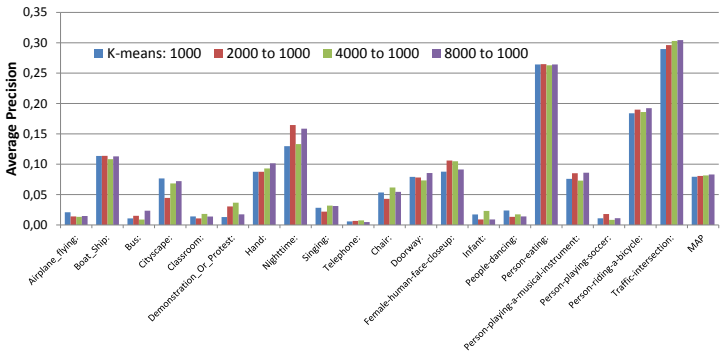
As the number of initial centers increases the individual concept dependent entropy minimization criteria for mapping suffers from overfitting as it generates dictionaries for each concept independently. The image level labeling does not translate well to increase the discriminative power of the BoW model built for each concept separately. This effect is carried on to the average (of concept

dependent) entropy minimization as the retrieval performance is adversely affected with the increase in the number of initial cluster centers in the first step of clustering.

Contrarily, merging neighboring clusters using minimization of the entropy of concept distribution given clustering improves retrieval performance with the increase in the size of the initial number of centers. Thus this merging criterion is used for evaluating the retrieval performance of 1000 word dictionaries obtained from larger dictionaries and we see improved performances as the initial dictionary size increases. Figure 2 shows concept-wise Average Precision (AP) results along with the MAP for the two baselines of 500 and 1000 visual words and entropy minimization based mappings with the value of p selected from 2, 4 and 8. Concepts like *Airplane Flying* and *Person Playing Soccer* that have a very low number of positives in the training set are adversly affected in their performance as the number of initial centers is increased. The MAP for 1000-words dictionary increases from 0.0796 (baseline) to 0.0831 with 8000 initial centers.



(a)



(b)

Fig. 2: Supervised clustering scores for 20 concepts using the first (best) entropy minimization criterion for (a) 500 and (b) 1000 visual words dictionaries

3.2 Alternative Mappings

We test retrieval performance for three simple modifications in the clustering criterion. To see the effect of including the unlabeled examples in the second step of our clustering framework we include all keypoints coming from the unlabeled keyframes as a new concept making the total number of concepts $m + 1$. The rest of the method remains the same which is the best performing entropy minimization criterion from the previous sub-section.

The second alternative is the relaxation of the constraint that only neighboring clusters can be merged, and the third version includes the unlabeled features while merging clusters over the whole clustering space. The MAP scores for a final dictionary of 500 words are presented in Table 2. Here we see good performance for the merging of 1000 initial clusters into 500 for the three alternatives and then a decline in the performance as the number of initial clusters is increased. This may be due to some overtraining as the number of choices for merging clusters are increased with the relaxation of constraints. However the performance is still better than the baseline results in each case.

Table 2: Mean Average Precision for 20 concepts using three alternatives of the concept distribution entropy minimization based mapping

Dictionary Size	**Unlab**	**All Space**	**Unlab All Space**
1000 to 500	0.0803	0.0805	0.0794
2000 to 500	0.0788	0.0795	0.0795
4000 to 500	0.0777	0.0755	0.0780

Finally we give upper bounds of improvement for each concept with highest average precision score selected from the retrieval results of different initial dictionary sizes for concept distribution entropy minimization with (i) neighboring constraint, (ii) inclusion of unlabeled features, (iii) relaxing the neighbor constraint and (iv) the inclusion of unlabeled features with the relaxation of neighbor constraint. Tables 3-(a) and 3-(b) shows the upper bounds of improvements for the two step clustering performed with its alternatives showing significant increase in the individual score for each concept for 500 and 1000 word final dictionary sizes. Individual scores for each concept improve significantly with only the concepts *Person eating* and *Traffic intersection* showing little improvement as their performance is already quite high. An exception is the concept *Airplane flying* which shows a decrease in performance with supervised merging for 1000 words dictionaries as shown in Table 3-(b).

3.3 Small and Informative vs Large Dictionaries

In the previous subsections we have shown results of our technique for building a visual dictionary and compared them to retrieval results of the baseline using

Table 3: Upperbounds of concept-wise improvements for dictionaries of (a) 500 visual words and (b) 1000 visual words

(a)

Semantic Concept	k-means	Entropy	Unlabeled	All-space	Unlab All	improvment
Airplane flying	0.0126	0.0257	**0.0266**	0.0266	0.0212	111%
Boat/Ship	0.0709	0.1005	0.1041	**0.1149**	0.1015	62%
Bus	0.0123	0.0170	0.0140	**0.0248**	0.0168	103%
Cityscape	0.0564	**0.0661**	0.0522	0.0652	0.0660	17%
Classroom	0.0132	0.0211	0.0265	**0.0463**	0.0197	250%
Demonstration	0.0158	0.0249	0.0239	**0.0325**	0.0168	106%
Hand	0.0809	0.0938	**0.0962**	0.0879	0.0872	19%
Nighttime	0.1037	0.1430	0.1487	**0.1545**	0.1460	49%
Singing	0.0415	0.0423	0.0399	0.0367	**0.0435**	5%
Telephone	0.0055	0.0072	0.0072	**0.0074**	0.0071	34%
Chair	0.0476	0.0581	**0.0618**	0.0576	0.0553	30%
Doorway	0.0865	**0.0915**	0.0857	0.0840	0.0810	6%
Female face closeup	0.0809	**0.1045**	0.1040	0.0959	0.0984	29%
Infant	0.0110	**0.0233**	0.0131	0.0170	0.0158	111%
People dancing	0.0116	0.0219	0.0218	**0.0301**	0.0196	159%
Person eating	0.2612	0.2664	0.2653	**0.2714**	0.2683	4%
Playing music	0.0779	0.0945	0.0926	**0.0949**	0.0884	22%
Person playing soccer	0.0081	**0.0203**	0.0179	0.0137	0.0138	151%
Person riding bicycle	0.1820	0.1926	0.1980	0.1949	**0.2100**	15%
Traffic intersection	0.2971	0.3036	0.3065	0.2947	**0.3096**	4%
MAP	**0.0739**	**0.0859**	**0.0853**	**0.0875**	**0.0843**	

(b)

Semantic Concept	k-means	Entropy	Unlabeled	All-space	Unlab All	improvment
Airplane flying	**0.0208**	0.0147	0.0130	0.0197	0.0178	-5%
Boat/Ship	0.1136	0.1129	0.1016	**0.1167**	0.1080	3%
Bus	0.0107	0.0235	0.0194	0.0212	**0.0317**	198%
Cityscape	0.0765	0.0720	**0.1042**	0.0697	0.0591	36%
Classroom	0.0140	0.0181	0.0177	**0.0394**	0.0197	181%
Demonstration	0.0130	**0.0366**	0.0339	0.0335	0.0332	181%
Hand	0.0876	0.1014	0.0997	**0.1023**	0.0959	17%
Nighttime	0.1297	0.1585	**0.1912**	0.1712	0.1570	47%
Singing	0.0282	0.0317	0.0397	**0.0405**	0.0367	44%
Telephone	0.0058	**0.0074**	0.0066	0.0059	0.0063	29%
Chair	0.0534	**0.0616**	0.0571	0.0581	0.0579	15%
Doorway	0.0792	**0.0855**	0.0825	0.0757	0.0777	8%
Female face closeup	0.0877	**0.1049**	0.1038	0.0957	0.1014	20%
Infant	0.0173	**0.0231**	0.0107	0.0151	0.0139	33%
People dancing	0.0238	0.0174	**0.0295**	0.0191	0.0181	24%
Person eating	0.2643	0.2643	0.2668	**0.2694**	0.2657	2%
Person playing music	0.0758	0.0860	**0.0890**	0.0830	0.0790	17%
Person playing soccer	0.0109	0.0111	**0.0129**	0.0120	0.0119	18%
Person riding bicycle	0.1839	0.1923	0.1875	0.1912	**0.2069**	13%
Traffic intersection	0.2967	**0.3045**	0.3014	0.2978	0.3024	3%
MAP	**0.0796**	**0.0864**	**0.0884**	**0.0869**	**0.0850**	

a dictionary obtained through sufficient number of k-means iterations. In those cases the sizes of the baseline and supervised dictionaries were same (500 words and 1000 words). We claimed that the retrieval performance of using dictionary obtained through supervised merging matches that of using a larger dictionary which is evident from the results in the Table 4.

Table 4: Comparing MAP for 20 concepts using three large dictionaries vs corresponding smaller supervised dictionaries of (a) 500 and (b) 1000 visual words

(a)

1000		**2000**		**4000**	
k-means	**500**	**k-means**	**500**	**k-means**	**500**
0.0796	0.0795	0.0814	0.0801	0.0830	0.0813

(b)

2000		**4000**		**8000**	
k-means	**1000**	**k-means**	**1000**	**k-means**	**1000**
0.0814	0.0806	0.0830	0.0816	0.0847	0.0831

SVM classifiers were trained for larger dictionaries containing 1000, 2000, 4000 and 8000 visual words. These are the dictionaries obtained in the first stage of the supervised merging using k-means. The training time for these larger BoWs is much higher and the performance is comparable to the smaller supervised BoW models. For example the 8 times smaller dictionary only results in 2% performance decrease as can be seen in the last two columns of the Table 4-(a) and less than 2% performance decrease as evident in the last two columns of the Table 4-(b).

The difference in performance will increase as the size of the first step dictionary increases as it becomes richer and richer. While merging helps to capture important semantic information the performance of the resulting supervised dictionary will be limited as the smaller dictionary will always be a coarser representation of the visual space.

As far as the computation overhead for supervised clustering is concerned it only uses the information from the image labels. Thus it does not perform any direct computation on the image features. The time complexity of supervised clustering is $O(n^3)$, with $n = p*D$, which is the cost borne once at the clustering stage during training phase giving a mapping from $p * D$ visual words to D visual words. After the two step clustering the costs of baseline and supervised dictionary for the remaining stages of video retrieval are similar.

4 Conclusions

We have seen that the discriminative ability of the Bag of Words model increases when performing the two step supervised clustering. The performance of a much smaller dictionary obtained through supervised merging reaches that of larger dictionaries obtained through k-means. The merging step is fast and incorporates already available label knowledge for calculation of the entropy. Although class specific merging of clusters overfits the BoW representation the performance is high as long as the initial number of clusters is kept low.

References

1. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027–1035. SODA '07, Philadelphia, PA, USA (2007), http://portal.acm.org/citation.cfm?id=1283383.1283494
2. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27 (2011), http://www.csie.ntu.edu.tw/ cjlin/libsvm
3. Hao, J., Jie, X.: Improved bags-of-words algorithm for scene recognition. In: Signal Processing Systems (ICSPS), 2010 2nd International Conference on. vol. 2, pp. V2–279 –V2–282 (2010)
4. Lin, C., Li, S., Su, S.: Image classification using adapted codebook. In: ITIME '09. vol. 1, pp. 1307 –1312 (2009)
5. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60, 91–110 (2004), http://portal.acm.org/citation.cfm?id=993451.996342
6. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: NIPS (2006), http://lear.inrialpes.fr/pubs/2006/MTJ06
7. Perronnin, F., Dance, C.R., Csurka, G., Bressan, M.: Adapted vocabularies for generic visual categorization. In: ECCV (4). pp. 464–475 (2006), http://dx.doi.org/10.1007/1174408536
8. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR '06. pp. 321–330 (2006), http://doi.acm.org/10.1145/1178677.1178722
9. Wang, L.: Toward a discriminative codebook: Codeword selection across multi-resolution. In: CVPR (2007), http://dx.doi.org/10.1109/CVPR.2007.383374
10. Winn, J., Criminisi, A., Minka, T.: Object categorization by learned universal visual dictionary. In: ICCV '05. pp. 1800–1807 (2005)