Enhancing image retrieval by an Exploration-Exploitation approach

Luca Piras[1], Giorgio Giacinto[1], and Roberto Paredes[2]

[1] Department of Electrical and Electronic Engineering University of Cagliari
Piazza D'armi, 09123 Cagliari (Italy)
{luca.piras, giacinto}@diee.unica.it
[2] Universidad Politécnica de Valencia
Camino de Vera s/n, Edif. 8G Acc. B, 46022 Valencia (Spain)
rparedes@dsic.upv.es

Abstract. In this paper, the Relevance Feedback procedure for Content Based Image Retrieval is considered as an Exploration-Exploitation approach. The proposed method exploits the information obtained from the relevance score as computed by a Nearest Neighbor approach in the *exploitation* step. The idea behind the Nearest Neighbor relevance feedback is to retrieve the immediate neighborhood of the area of the feature space where relevant images are found. The exploitation step aims at returning to the user the maximum number of relevant images in a local region of the feature space. On the other hand, the *exploration* step aims at driving the search towards different areas of the feature space in order to discover not only relevant images but also informative images. Similar ideas have been proposed with Support Vector Machines, where the choice of the informative images has been driven by the closeness to the decision boundary. Here, we propose a rather simple method to explore the representation space in order to present to the user a wider variety of images. Reported results show that the proposed technique allows to improve the performance in terms of average precision and that the improvements are higher if compared to techniques that use an SVM approach.

Keywords: Algorithms, Active Learning, max-min

1 Introduction

Nowadays, the possibility for people to easily create, store, and share, vast amount of multimedia documents is a problem that the pattern recognition community is facing since several years. Digital cameras allow capturing an unlimited number of photos and videos, thanks to the fact that they are also embedded in a number of portable devices. This vast amount of media archives needs to be organized in order to ease future search tasks. It is easy to see that it is often impractical to automatically label the content of each image or different portions of videos by recognizing the objects in the scene, as we should have templates for each object in different positions, lighting conditions, occlusions, etc. [26].

It is quite easy to see that each picture and video may be characterized by a large number of concepts depending on the level of detail used to describe the scene, or the focus in the description. Moreover, different users may describe an image using different categories, and the same user may classify the same image in different ways depending on the context. Sometimes, an image may contain one or more concepts that can be prevalent with respect to others, so that if a large number of people are asked to label the image, they may unanimously use the same label. Nevertheless, it is worth noting that a concept may be also decomposed in a number of "elementar" concepts. For example, an image of a car can have additional concepts, like the color of the car, the presence of humans or objects, etc. Thus, for a given image or video-shot, the same user may focus on different aspects. How the task of retrieving similar images or videos from an archive can be solved by automatic procedures? How can we design procedures that automatically tune the similarity measure to adapt to the visual concept the user is looking for? Once again, the target of the classification problem cannot be clearly defined beforehand but must be designed to explicitly take into account user needs [10].

In the field of content based multimedia retrieval, a number of review papers pointed out the difficulties in providing effective similarity measure that can cope with the broad domain of content of multimedia archives [6, 21, 28]. The shortcomings of current techniques developed for image and video has been clearly shown by Pavlidis [25]. While systems tailored to a particular image domain (e.g., medical images) can exhibit quite impressive performances, the use of these systems on unconstrained domains reveals their inability to adapt dynamically to new concepts [27]. One solution is to have the user manually label a small set of representative images as relevant or non relevant to the query (the so-called relevance feedback) that are used as training set for updating the similarity measure [33]. In this system is not uncommon that, after the first feedback iterations, the number of relevant images retrieved increases quickly. However, the system typically stops providing new relevant images despite of the user interaction. The reason lies in the way in which images are presented to the user. In fact, usually the best ranked images are retrieved after each round of feedback, and these images are usually retrieved in a small local area of the feature space. As a consequence, the search often converges towards a local optimum, without taking into account images in other areas of the feature space. In order to address this kind of problems, we propose an *Exploitation-Exploration* mechanism where the exploration step is inspired by *Active Learning* [4]. Our approach requires the system to choose not only the most relevant images according to the user judgement, but also the most informative images that allows driving the search in more promising regions of the feature space. The key issue is how to choose the most informative images. Usually this approach has been used in systems based on discriminative functions, i.e. system that builds a decision function which classifies the unlabelled data.

One method to select informative images is based on choosing the patterns closest to the decision boundary, as described in [31, 16] where SVM based on active learning are used. In [3], the authors proposed to learn two SVM classifiers in two uncorrelated feature spaces as color and texture. The classifiers have been then used to classify the images and the unlabelled ones that received a different label in the two feature spaces, have been chosen to be shown to the user. In addition, different criteria have been proposed over the years as the minimization of expected average precision [13], or the maximization of the entropy [19]. In the latter paper the authors learned an SVM on the labelled images, mapped the SVM outputs into probabilities and chose the images with the probability to belong to relevant class nearest to 0.5. In [17] the authors instead to use the proximity to the theoretical decision boundary as measure of the information capability of the training images, propose a clarity index that takes into account the rank of each image with respect to those of the known relevant and non relevant images. The images with the lowest values of clarity are chosen as training images.

Conventional SVM active learning is designed to select a single example for each learning iteration but, as suggested in [15], usually in a relevance feedback iteration the user labels multiple image examples as being relevant or non relevant. In this case it is possible that the system selects similar images to learn the SVM. The authors, to address this problem, proposed a *Batch Mode Active Learning* technique that chooses the most suitable unlabelled examples one at a time. An interesting approach has been proposed in [34] where the authors propose a novel paradigm of active learning, which is able to estimate the probability density function (pdf) of the underlying query distribution to avoid the risk of learning on a completely unknown distribution. The estimated pdf, together with the distribution of the classifier outcomes, is used to guide the sampling, in the way that it is possible to give priority to two types of instances to label in the next iteration, namely instances in the area where the probability to find relevant pattern is high (for boosting the retrieval) and instances in the uncertain area (for figuring out the new decision boundary). In [22] the authors, instead of using SVM, proposed a selective sampling for Nearest Neighbor classifiers. In order to choose the most informative patterns they suggest to consider not only the uncertainty of the candidate sample point, but also the effect of its classification on the remaining unlabelled points. For this reason, their *lookahead algorithm for selective sampling* considers sampling sequences of neighboring patterns of length k, and selects an example that leads to the best sequence. The best sequence is the one whose samples have the highest conditional class probabilities. Also in [18] the authors proposed a probabilistic variant of the k-Nearest Neighbor method for active learning in multi-class scenarios. After that they defined a probability measure, based on the pairwise distances between data points, they used the Shannon entropy as "uncertain" measure over the class labels in order to maximize the discriminating capabilities of the model.

In this paper we consider the most informative images as those that are distributed around the images that have been labelled by the user along all the representation space. This task can be accomplished by resorting to an Active

Learning approach based on hierarchical clustering of the data [5]. Although this technique allows obtaining good results, it is quite computationally expensive. In order to use a computationally cheap method, we performed the exploration phase through a *max-min* approach that showed good results in similar tasks such as the initialization of the *c-means* algorithm [20].

This paper is organized as follows. Section 2 briefly reviews the Nearest Neighbor approach used in order to assign a Relevance Score to the images. Section 3 introduces the proposed technique and describes how exploit it in the Relevance Feedback iterations. Experimental results are reported in Section 4. Conclusions are drawn in Section 5.

2 Nearest-Neighbor Relevance Feedback for Relevance Score

The use of the Nearest-Neighbor paradigm has been inspired by classification techniques based on the "nearest case", which are used in pattern recognition and machine learning for classification and outlier detection. In addition, nearest-neighbor techniques have been also used in the context of "active learning", which is closely related to relevance feedback [22]. In particular, recent works on outlier detection and one-class classification clearly pointed out the effectiveness of nearest-neighbor approaches to identify objects belonging to the target class (i.e., the relevant images), while rejecting all other objects (i.e., non relevant images) [2, 30]. This approach is suited to cases when it is difficult to produce a high-level generalization of a "class" of objects.

This approach can be used for estimating image relevance in CBIR as each "relevant" image as well as each "non relevant" image can be considered as individual "cases" or "instances" against which the images of the database should be compared [12].

In this paper, a method proposed in [9] has been used, where a score is assigned to each image of a database according to its distance from the nearest image belonging to the target class, and the distance from the nearest image belonging to a different class. This score is further combined to a score related to the distance of the image from the region of relevant images. The combined score is computed as follows:

$$rel(\mathbf{x}) = \left(\frac{n/t}{1+n/t}\right) \cdot rel_{BQS}(\mathbf{x}) + \left(\frac{1}{1+n/t}\right) \cdot rel_{NN}(\mathbf{x}) \tag{1}$$

where n and t are the number of non-relevant images and the whole number of images retrieved after the last iteration, respectively. The two terms rel_{NN} and rel_{BQS} are computed as follows:

$$rel_{NN}(\mathbf{x}) = \frac{\|\mathbf{x} - NN^{nr}(\mathbf{x})\|}{\|\mathbf{x} - NN^{r}(\mathbf{x})\| + \|\mathbf{x} - NN^{nr}(\mathbf{x})\|} \tag{2}$$

where $NN^r(\mathbf{x})$ and $NN^{nr}(\mathbf{x})$ denote the relevant and the non relevant Nearest Neighbor of $\mathbf{x}$, respectively, and $\|\cdot\|$ is the metric defined in the feature space at hand,

$$rel_{BQS}(\mathbf{x}) = \frac{1 - e^{1 - d_{BQS}(\mathbf{x}) \big/ \max_i d_{BQS}(\mathbf{x}_i)}}{1 - e} \tag{3}$$

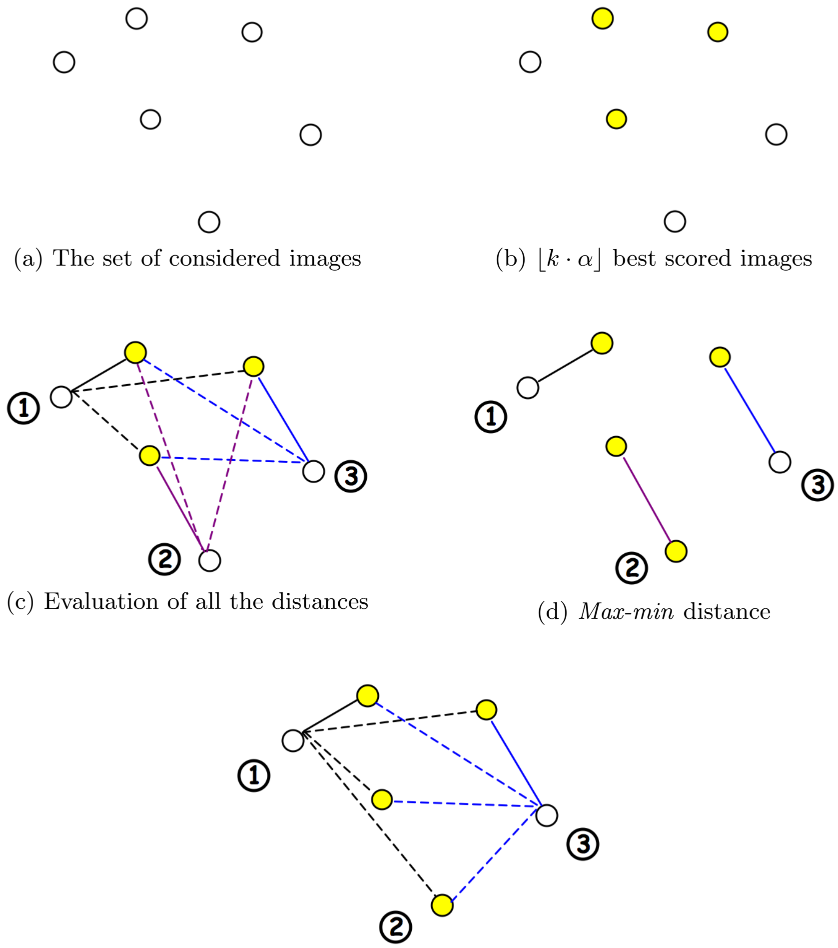
where e is the *Euler's number*, i is the index of all images in the database and d_{BQS} is the distance of image $\mathbf{x}$ from a reference vector computed according to the Bayes decision theory (Bayes Query Shifting, BQS) [11].

3 The exploration phase

Typically, an image retrieval system with relevance feedback works as follows: after the submission of a query, the system, according to a similarity measure, scores all images in the database and presents the k best scored ones to the user. One of the problems in this kind of behavior is that in the following iterations the search could be driven by the (probably few) relevant images retrieved so far, and the system could be trapped in a limited area of the feature space. Sometimes, neither the non relevant images, that are also considered in the evaluation of the relevance score, can help to get out of this situation. In fact, by always using the same set of relevant images iteration by iteration, the search can "take a wrong way". In order to face this problem, the proposed method selects within a certain number of best scored images, those images that are "not too close" to the "classical search area". The meaning of "not too close" and "classical search area" will be explained in the following.

Let us define k as the number of the images to return to the user and $\lfloor k \cdot \alpha \rfloor$ as the fixed number of the "seed images". Let us also define $\lfloor (k - \lfloor k \cdot \alpha \rfloor) \cdot \beta \rfloor$ as the number of images among which to choose the most informative ones. In the previous formulas, the parameter α can assume values between $\frac{1}{k}$ and 1, and $\beta \geq 1$. Summing up, with the *Exploration-Exploitation* approach k images are shown to the user, the best scored $\lfloor k \cdot \alpha \rfloor$ are selected beforehand, while the other $(k - \lfloor k \cdot \alpha \rfloor)$ are chosen through a *max-min* approach between the $\lfloor (k - \lfloor k \cdot \alpha \rfloor) \cdot \beta \rfloor$ best scored images. It is clear that if $\alpha = \frac{1}{k}$ all the images, apart from the query, are selected in an "active" way, on the contrary, when it is equal to 1, they system shows to the user the best k scored images as in the classical Nearest Neighbor approach. The same happens when $\beta = 1$, as $(k - \lfloor k \cdot \alpha \rfloor)$ images will be chosen from a set of $(k - \lfloor k \cdot \alpha \rfloor)$ best scored images. The *max-min* approach selects an image from a set evaluating all the distances between the seed images and the images in the set and choosing for each of them the shortest. The images are then sorted according to these distances and that with the maximum distance it is selected. The idea behind the *max-min* technique is similar to that of pruning, that is, it pays attention to the relevant images that are most far apart from each other because they are more likely to be the most different from those that are usually used [20]. To better explain the algorithm, Figure 1 shows an example where $k = 5$, $\alpha = 0.6$, and $\beta = 1.5$.

(a) For each image in the database a score is computed according to Eq. (1);



(a) The set of considered images

(b) $\lfloor k \cdot \alpha \rfloor$ best scored images

(c) Evaluation of all the distances

(d) *Max-min* distance

(e) New round of selection

Fig. 1. Exploration-Exploitation algorithm.

- **(b)** the three best scored images are used as seeds of the search ($\lfloor 5 \cdot 0.6 \rfloor = 3$);
- **(c)** for each of the remaining images ($\lfloor (5 - \lfloor 5 \cdot 0.6 \rfloor) \cdot 1.5 \rfloor = 3$) the distances with the seed images are evaluated, and the minimum ones are chosen;
- **(d)** the image with the largest minimum distance (image **(2)**) is chosen to be added to the seed images.
- **(e)** in order to add the fifth image to be shown to the user ($k = 5$), the algorithm restarts from step **(c)**.

It is clear that if $\alpha = \frac{1}{k}$, all the images, apart from the query, are selected following the *Exploration* approach. On the other hand, when $\alpha = 1$ the best k scored images, as in the classical Nearest Neighbor approach, are shown to the user. The same happens when $\beta = 1$, in fact in this situation $(k - \lfloor k \cdot \alpha \rfloor)$ images from a set of $(k - \lfloor k \cdot \alpha \rfloor)$ best scored images are chosen.

4 Experimental Results

4.1 Datasets

Experiments have been carried out using three datasets, namely the Caltech-256 dataset, from the California Institute of Technology[3], the WANG dataset[4], and the Microsoft Research Cambridge Object Recognition Image Database[5] (in the following referred to as MSRC). The first dataset consists of 30607 images subdivided into 257 semantic classes [14], the WANG dataset consists of a subset of 1000 images of the Corel stock photo database which have been manually selected and which form 10 classes of 100 images each [32], and MSRC contains 4320 images subdivided into 17 "main" classes, each of which is further subdivided into subclasses, for a total of 33 semantic classes [35]. From Caltech-256, the *Edge Histogram* descriptor [1] has been extracted using the open source library LIRE (Lucene Image REtrieval) [24]. The images from the WANG dataset are represented by a 512-dimensional colour histogram and a 512-dimensional Tamura texture feature histogram [29] concatenated in a single vector [7]. The images of MSRC are represented by a vector of 4096 components of SIFT descriptors [23] extracted at Harris interest points [8, 7]. The WANG, MSRC and Caltech-256 datasets represents image retrieval tasks of different complexity. In fact, the WANG dataset is usually considered an easy task in the Image Retrieval context. The MSRC dataset is mainly used in the object recognition domain, as the pictures usually contain one object "captured" from a particular point of view (front, side, rear, etc.) or at most more objects of the same type. The Caltech-256 dataset is also widely used in object recognition, and it can be considered a more difficult task than the one represented in the MSRC dataset: the semantic concepts in the Caltech dataset are more loosely related to the image content. As example, the class "marriages" contains images of newlyweds as well as images of wedding cakes.

4.2 Experimental Setup

In order to test the performances of the proposed approaches, 500 query images from Caltech-256 dataset have been randomly extracted, so that they cover all the semantic classes. For the WANG and MSRC datasets, each image is used as query. Relevance feedback is performed by marking images belonging to the same class of the query as relevant, and all other images in the pool of k to-be-labelled images as non-relevant. Performance is evaluated in terms of mean average precision taking into account all the relevant images as they are ordered

[3] http://www.vision.caltech.edu/Image_Datasets/Caltech256/
[4] http://wang.ist.psu.edu/docs/related.shtml
[5] http://research.microsoft.com/downloads

by the classifier. In order to choose the most suitable values of the parameters α and β, a number of preliminary experiments have been performed. Accordingly, it has been fixed the value of β at 10, whereas different values of α have tested, and the related results are reported for comparison purposes. The relevance score at each image has been assigned according to Eq. (1), and in the graph it has been referred to as the *Relevance Score*, whereas the case $\alpha = \frac{1}{k}$ has been referred to as $\alpha = 0\%$. For comparison purposes, relevance feedback has been also computed by a "pure" SVM classifier with an RBF kernel and an SVM_{ACTIVE} as well. The Active Learning has been performed according to the approach proposed in [17].

4.3 Results

Figures 2(a), 2(b), and 2(c) show the average precision evaluated using the WANG, the MSRC, and the Caltech datasets, respectively. Observing the Figure 2 it is quite clear to see how the lines have a quite different trend w.r.t. the value of α. In fact, the lower the value of α in Figure 2, the better the measured average precision. The only exception in the above trend can be seen in the case of $\alpha = \frac{1}{k}$ (referred to as $\alpha = 0\%$) in Figure 2(a) and (c), where the average precision is lower than the values obtained with $\alpha = 0.25\%$. The reason of this behavior probably is due to the fact that only the query is not enough as "seed image" in order to begin the search. In the case of the WANG and MSRC dataset it is also easy to see how the lower improvement obtained using the Exploration-Exploitation approach ($\alpha = 0.75\%$) w.r.t. the method without "Exploration" phase, overcomes the improvement of the SVM_{ACTIVE} w.r.t the "pure" SVM.

5 Conclusion

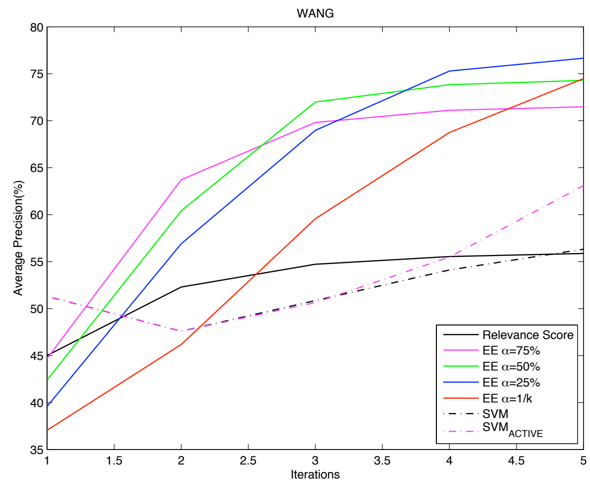
In this paper the problem of low informative training sets has been faced exploiting the Nearest Neighbor paradigm in an Exploration-Exploitation task. The proposed algorithm subdivides the NN approach in two phases: *Exploitation* and *Exploration*. In the Exploitation phase a relevance score is assigned to each image in the database, while in the Exploration phase a certain number of best scored images is drawn, and by means of a max-min approach based on the distances among the images, the images that will be shown to the user to capture her feedback are chosen. The obtained results clearly show that the bigger the percentage of the images, proposed to the user, the larger is the obtained average precision. According to these results, it is possible to say that the proposed Exploration-Exploitation approach is able to move the search in areas of the feature space usually "unexplored".

References

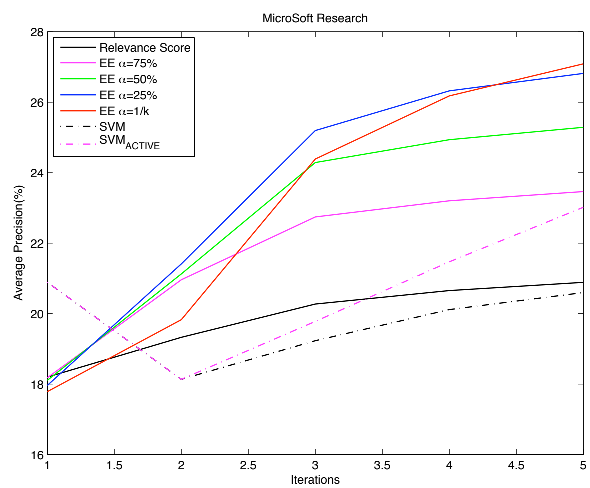
1. Information technology - Multimedia content description interface - Part 3: Visual, ISO/IEC Std. 15938-3:2003 (2003)

2. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying density-based local outliers. In: Chen, W., Naughton, J.F., Bernstein, P.A. (eds.) SIGMOD Conference. pp. 93–104. ACM (2000)
3. Cheng, J., Wang, K.: Active learning for image retrieval with co-svm. Pattern Recognition 40(1), 330–334 (2007)
4. Cohn, D.A., Atlas, L.E., Ladner, R.E.: Improving generalization with active learning. Machine Learning 15(2), 201–221 (1994)
5. Dasgupta, S., Hsu, D.: Hierarchical sampling for active learning. In: Cohen, W.W., McCallum, A., Roweis, S.T. (eds.) ICML. ACM International Conference Proceeding Series, vol. 307, pp. 208–215. ACM (2008)
6. Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys 40(2), 1–60 (2008)
7. Deselaers, T., Keysers, D., Ney, H.: Features for image retrieval: an experimental comparison. Inf. Retr. 11(2), 77–107 (2008)
8. Dorkó, G.: Selection of Discriminative Regions and Local Descriptors for Generic Object Class Recognition. Ph.D. thesis, Institut National Polytechnique de Grenoble (2006)
9. Giacinto, G.: A nearest-neighbor approach to relevance feedback in content based image retrieval. In: CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval. pp. 456–463. ACM, New York, NY, USA (2007)
10. Giacinto, G.: Moving targets in computer security and multimedia retrieval. Trans. MLDM 4(1), 30–52 (2011)
11. Giacinto, G., Roli, F.: Bayesian relevance feedback for content-based image retrieval. Pattern Recognition 37(7), 1499–1508 (2004)
12. Giacinto, G., Roli, F.: Instance-based relevance feedback for image retrieval. In: Saul, L.K., Weiss, Y., Bottou, L. (eds.) Advances in Neural Information Processing Systems 17, pp. 489–496. MIT Press (2005)
13. Gosselin, P.H., Cord, M.: Active learning methods for interactive image retrieval. IEEE Transactions on Image Processing 17(7), 1200–1211 (2008)
14. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology (2007), `http://authors.library.caltech.edu/7694`
15. Hoi, S.C.H., Jin, R., Zhu, J., Lyu, M.R.: Semisupervised svm batch mode active learning with applications to image retrieval. ACM Trans. Inf. Syst. 27(3), 16:1–16:29 (2009)
16. Hoi, S.C.H., Lyu, M.R.: A semi-supervised active learning framework for image retrieval. In: CVPR (2). pp. 302–309. IEEE Computer Society (2005)
17. Huang, T., Dagli, C., Rajaram, S., Chang, E., Mandel, M., Poliner, G., Ellis, D.: Active learning for interactive multimedia retrieval. Proceedings of the IEEE 96(4), 648 –667 (april 2008)
18. Jain, P., Kapoor, A.: Active learning for large multi-class problems. In: CVPR. pp. 762–769. IEEE (2009)
19. Jing, F., Li, M., Zhang, H., Zhang, B.: Entropy-based active learning with support vector machines for content-based image retrieval. In: ICME. pp. 85–88. IEEE (2004)
20. Katsavounidis, I., Jay Kuo, C.C., Zhang, Z.: A new initialization technique for generalized lloyd iteration. Signal Processing Letters, IEEE 1(10), 144 –146 (oct 1994)
21. Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Trans. Multimedia Comput. Commun. Appl. 2(1), 1–19 (2006)

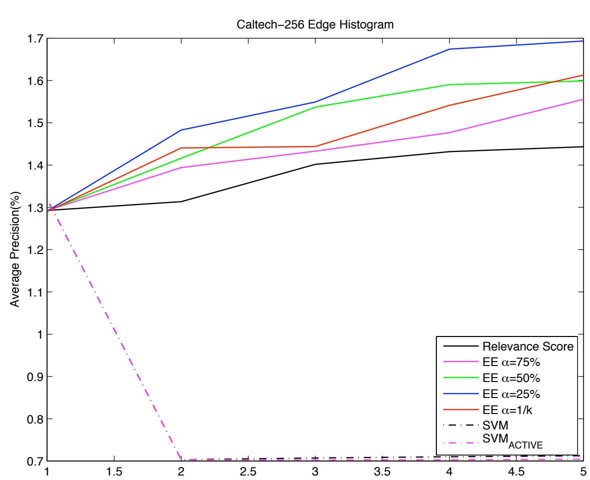
22. Lindenbaum, M., Markovitch, S., Rusakov, D.: Selective sampling for nearest neighbor classifiers. Machine Learning 54(2), 125–152 (2004)
23. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
24. Lux, M., Chatzichristofis, S.A.: Lire: lucene image retrieval: an extensible java cbir library. In: MM '08: Proceeding of the 16th ACM international conference on Multimedia. pp. 1085–1088. ACM, New York, NY, USA (2008)
25. Pavlidis, T.: Limitations of content-based image retrieval. http://theopavlidis.com/technology/CBIR/PaperB/vers3.htm (2008)
26. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: Labelme: A database and web-based tool for image annotation. International Journal of Computer Vision 77(1-3), 157–173 (2008)
27. Sivic, J., Zisserman, A.: Efficient visual search for objects in videos. Proceedings of the IEEE 96(4), 548 –566 (april 2008)
28. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. 22(12), 1349–1380 (2000)
29. Tamura, H., Mori, S., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Trans. Systems, Man and Cybernetics 8(6), 460–473 (June 1978)
30. Tax, D.M.: One-class classification. Ph.D. thesis, Delft University of Technology, Delft, The Netherlands (June 2001)
31. Tong, S., Chang, E.Y.: Support vector machine active learning for image retrieval. In: ACM Multimedia. pp. 107–118 (2001)
32. Wang, J.Z., Li, J., Wiederhold, G.: Simplicity: Semantics-sensitive integrated matching for picture libraries. IEEE Trans. Pattern Anal. Mach. Intell. 23(9), 947–963 (2001)
33. Wang, J., Hua, X.S.: Interactive image search by color map. ACM TIST 3(1), 12 (2011)
34. Wei, X.Y., Yang, Z.Q.: Coached active learning for interactive video search. In: Candan, K.S., Panchanathan, S., Prabhakaran, B., Sundaram, H., chi Feng, W., Sebe, N. (eds.) ACM Multimedia. pp. 443–452. ACM (2011)
35. Winn, J.M., Criminisi, A., Minka, T.P.: Object categorization by learned universal visual dictionary. In: ICCV. pp. 1800–1807. IEEE Computer Society (2005)



(a) WANG Dataset



(b) MSRC Dataset



(c) Caltech-256 Dataset

Fig. 2. Average Precision for 5 rounds of relevance feedback.