Passive-Aggressive Online Learning for Relevance Feedback in Content Based Image Retrieval

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.*
[2]*Universidad Politécnica de Valencia Camino de Vera s/n, Edif. 8G Acc. B, 46022 Valencia, Spain*
{*luca.piras, giacinto*}*@diee.unica.it, rparedes@dsic.upv.es*

Keywords: Online Learning: Image Retrieval: Relevance Feedback.

Abstract: The increasing availability of large archives of digital images has pushed the need for effective image retrieval systems. Relevance Feedback (RF) techniques, where the user is involved in an iterative process to refine the search, have been recently formulated in terms of classification paradigms in low-level feature spaces. Two main issues arises in this formulation, namely the small size of the training set, and the unbalance between the class of relevant images and all other non-relevant images. To address these issues, in this paper we propose to formulate the RF paradigm in terms of Passive-Aggressive on-line learning approaches. These approaches are particularly suited to be implemented in RF because of their iterative nature, which allows further improvements in the image search process. The reported results show that the performances attained by the proposed algorithm are comparable, and in many cases higher, than those attained by other RF approaches.

1 INTRODUCTION

Historically to incorporate pattern recognition into image retrieval, we distinguish different cases. In some cases, a Content Based Image Retrieval (CBIR) system is applied for a special task, where the images searched are from a particular domain and there is a set of queries and relevant images known. In these cases, it is possible to learn parameters for a system to optimise retrieval performance. The other and more generally applicable cases where pattern recognition techniques can be used in CBIR involves relevance feedback. The vast amount of digital pictures produced, stored, and shared daily, demands effective tools for finding similar images. Although CBIR systems have been under investigation since the 90's (Nie et al., 2012), still the most popular way of querying image archives is through the use of textual tags or captions associated to images. On the other hand, accessing images by textual information is not always satisfactory for many users' needs, because a large number of concepts a user is interested in can be better expressed through a query image rather than by a text description. In the last years, begun to emerge systems that retrieve images also through their content as, for example, Google Similar Images that combines information from tags, comments, and user's clicks with content based features. Despite the good results that can be achieved with this type of systems the computational cost required for this type of image indexing makes this approach difficult to integrate into less complex systems. For this reason, research on systems completely content based and not so computationally expensive is still very active.

The problem of finding images according to the user's requirements can be divided into a two-step procedure. First, the user submits a query image containing the concepts of interest. The system assigns to each image in the database a relevance score related to the similarity between the images and the query, according to some suitable similarity measure. Then, a number of best scored images are returned to the user who labels them as being relevant or not according to the concept in mind. These images are then used by the system to adapt the search in order to provide the user with new images.

One of the first approaches proposed in the literature for finding relevant images according to user's feedback, is based on techniques developed in the context of the *Information Retrieval* field (Zhou and Huang, 2003). These techniques are based on the concept that there is a region in the feature space where relevant images are clustered. Thus, one approach to find relevant images is to move the query towards the area that should be more densely populated with relevant images (Zhou and Huang, 2003). Differently

from the above methods that are essentially based on density estimation, there is another family of relevance feedback techniques based on discriminative learning, i.e. methods that learn from a set of images labelled as being relevant or not. In this family, a leading role is played by Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000). The idea behind the SVM is to find a hyperplane in the feature space that divides it into two subspaces. The first populated by relevant samples, the second one by non-relevant samples (Zhou and Huang, 2003).

Other approaches to exploit relevance feedback are aimed at computing a distance metric such that the distance in the low-level feature space is consistent with the users' relevance judgements. This metric minimizes the distance between similar images, and meanwhile maximizes the distance between the feature vectors of dissimilar images (Deselaers et al., 2008). While the above approaches are based on a batch (off-line) formulation of the learning process, the iterative nature of relevance feedback makes it suited to on-line learning approaches. Moreover, as the user's intention could change along the RF process, an on-line learning formulation allows the system to adapt the model accordingly.

Online learning techniques address a number of classification problems, such as binary and multiclass categorization, as well as regression and sequence prediction problems. Typically, on-line learning algorithms evaluate one pattern at time, and, for each pattern, the algorithm predicts if it belongs or not to the class of interest. Each pattern is associated with a unique label $y_t \in \{+1, -1\}$ where $\{+1\}$ indicates that the pattern belongs to that class, while $\{-1\}$ indicates that the pattern does not belong to that class. After each class assignment, the predicted class is compared to the true class. According to the outcome of the previous prediction the algorithm adapts its prediction rule for the next evaluation in order to improve the performances. Among the different online learning methods, we focus our attention to the Passive-Aggressive (PA) family of algorithms (Crammer et al., 2006), as they are based on the margin paradigm used by SVM. This method has been successfully used for ranking images after a text query (Grangier and Bengio, 2008), and for video tagging applications (Paredes et al., 2009) among others. In this paper, we propose the use of the PA paradigm to exploit Relevance Feedback (RF) in Content Based Image Retrieval (CBIR). We will show that the proposed on-line approach allows attaining a very fast and compact RF method, a property that is very important in a real-world CBIR scenario.

This paper is organized as follows. Section 2 presents the proposed technique based on a linear model, while Section 3 present the kernel version. Experimental results reported in Section 4 shows that a passive-aggressive approach provides good retrieval performances when compared to approaches based on classical classification approaches. Conclusions are drawn in Section 5.

2 ONLINE LEARNING FOR RELEVANCE FEEDBACK IN CBIR

Relevance Feedback is an interactive process where, given an initial query, the user is presented with the top-ranked images from the database, and is asked to mark them as being relevant or not. Then, these judgements are sent back to the retrieval system that exploits user's feedback to refine the search function, and provide the user with a new set of (hopefully) better results. In each iteration of relevance feedback, the system can use learning techniques to refine the similarity score measure, and thus to improve the results.

Online learning algorithms have been proposed for classification problems, where the goal, in the binary case, is to assign a pattern to one of the two classes. In image retrieval settings, the goal is to rank images according to the similarity to the query. Thus, online learning approaches need to be adapted. The approach proposed in this paper is inspired by the work of (Grangier and Bengio, 2008), that adapted Passive-Aggressive (PA) online classification approaches to retrieval scenarios. In the following, we will briefly summarize the approach.

2.1 A Linear Model for Content Based Image Retrieval

In the case of on-line binary classification problems, one of the simplest classification approach is the linear classifier, where the classification function is actually implemented by a vector of weights. Given a vector $\mathbf{w}$, and a pattern $\mathbf{x}$, the sign of the product $(\mathbf{w} \cdot \mathbf{x})$ can be used to classify the pattern as belonging to one or the two classes, while the norm of the product $|\mathbf{w} \cdot \mathbf{x}|$ is the degree of confidence in the classification. We can modify the above definition of linear classifier so that the result can be interpreted as a relevance score assigned to image $\mathbf{x}$:

$$score(\mathbf{x}_i) = \mathbf{w} \cdot \mathbf{x}_i, \quad (1)$$

where $\mathbf{w}$ is the weight vector to be learnt. According to (Grangier and Bengio, 2008), If the above men-

tioned score is used to rank images, the desired behavior of a CBIR system can be formulated as follows:

$$\forall \mathbf{x}_r \in R, \forall \mathbf{x}_n \in N, \mathbf{w} \cdot \mathbf{x}_r > \mathbf{w} \cdot \mathbf{x}_n, \tag{2}$$

where R and N represent relevant and non relevant images, respectively. In order to attain this behavior, we can maximize the following expression:

$$\sum_{\forall \mathbf{x}_r \in R} \sum_{\forall \mathbf{x}_n \in N} \mathbf{w} (\mathbf{x}_r - \mathbf{x}_n). \tag{3}$$

Several methods can be used in order to compute $\mathbf{w}$ that maximizes Eq. (3). Here we follow an idea similar to the one proposed in (Grangier and Bengio, 2008) based on PA online learning (Crammer et al., 2006) to compute the weight $\mathbf{w}$ by solving the following equation:

$$\mathbf{w}_t = \underset{\mathbf{w}_{t+1}}{\operatorname{argmin}} \frac{1}{2} ||\mathbf{w}_{t+1} - \mathbf{w}_t||^2 + C l_t, \tag{4}$$

with

$$l_t = \begin{cases} 0 & \text{if } \mathbf{w}_t (\mathbf{x}_r - \mathbf{x}_n) > 1 \\ 1 - \mathbf{w}_t (\mathbf{x}_r - \mathbf{x}_n) & \text{otherwise.} \end{cases} \tag{5}$$

where t is the iteration and l_t is the loss function at iteration t. The first term in Eq. (4), $\frac{1}{2}||\mathbf{w}_{t+1} - \mathbf{w}_t||^2$, plays the *passive* role in the algorithms as it forces the values of the weights obtained at consecutive iterations to be close to each other, thus taking into account all the information learned in the past iterations. Eq. (4) also shows that the second term, i.e., the loss function that pays the role of the *aggressive* term, is "smoothed" by the parameter C that avoids an excessive fluctuation of the weight between two iterations in succession. According to (Crammer et al., 2006) the solution of Eq. (4) is given by:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \Gamma_t (\mathbf{x}_r - \mathbf{x}_n), \tag{6}$$

where

$$\Gamma_t = \min \left\{ C, \frac{l_t}{||\mathbf{x}_r - \mathbf{x}_n||^2} \right\}. \tag{7}$$

2.2 Online Learning Algorithm for Relevance Feedback

The PA learning framework proposed in the previous section for image retrieval, can be used to implement relevance feedback according to the following procedure:

i) The user provides a query image to the system, the system computes the similarity of the query with all the images in the database, and images are then sorted in order of decreasing similarity with the query. The first k images are presented to the user who labels them as being relevant or not;

ii) henceforth, the on-line learning algorithm begins. Each component of the initial weight vector $\mathbf{w}_0$ is initialized to 0;

iii) a pair of images $\mathbf{x}_r$ and $\mathbf{x}_n$ is randomly drawn from the set of relevant images (R), and the set of non-relevant images (N), respectively. The random drawing is repeated for a fixed number of rounds, and at each round the weight vector is updated accordingly;

iv) using the updated weight vector, a new score for each image of the database is computed according to Eq. (1). The images are sorted according to the scores, and the first k are labelled by the user as in step **i)**;

v) a new relevance feedback iteration begins with a new iterative update of the weight vector $\mathbf{w}$ obtained from the previous iteration, until the user is satisfied.

The choice of using a predefined number of weight updates in step iii) allows to handle different proportions in relevant and non relevant images. If the dimension of the two sets of relevant and non relevant images is large, then the overall number of possible pairs is large, and the procedure would be computationally too expensive. On the other hand, if one of the two sets is much smaller than the other (as it typically happens in relevance feedback scenarios, where non relevant images typically outnumbers relevant images) the weight would be unbalanced toward the bigger set. It turns out that, by using a fixed number of rounds, images belonging to the smaller group are drawn more than once.

Finally, the motivation behind the use of random drawing is twofold. First of all, the user labels the images as relevant and non relevant to the query but does not provide any ranking, so by random drawing pairs of images, all the images can contribute to the evaluation of the weight. Second, the final result is different if the same image is considered more than once in different rounds of the weight update procedure. In fact the algorithm takes into account, thanks to the first term of Eq. (4), all the previous updates, and thus if the same images are considered two or three times in the same update, an improvement in the weight value can be attained in the same way as considering two or three different images.

3 KERNEL FORMULATION

In order to formulate the algorithm by resorting to the so called *kernel trick*, it is useful to use a classification model instead of a ranking model. Accord-

ingly, each image $\mathbf{x}_i$ is associated with a unique label $y_t \in \{+1,-1\}$ where $\{+1\}$ indicates the relevant images and $\{-1\}$ the non relevant ones. Moreover, the goal should not be formulated as the maximization of the distances between relevant and non relevant images, rather it should be formulated as the maximization of the *margin* of the decision. In order to reflect this change in perspective, we can modify the update function (6) as follows:

$$\mathbf{w}_t = \sum_{j=0}^{t-1} \Gamma_j \, y_j \, \mathbf{x}_j, \tag{8}$$

and therefore

$$score_t(\mathbf{x}_i) = \mathbf{w}_t \cdot \mathbf{x}_i = \sum_{j=0}^{t-1} \Gamma_j \, y_j \, (\mathbf{x}_j \cdot \mathbf{x}_i). \tag{9}$$

The inner product in the right hand side at Eq. (9) can be replaced by a kernel expression $K(\mathbf{x},\mathbf{x}')$ thus obtaining

$$score_t(\mathbf{x}_i) = \sum_{j=0}^{t-1} \Gamma_j \, y_j \, K(\mathbf{x}_j,\mathbf{x}_i), \tag{10}$$

where according to (Crammer et al., 2006):

$$\Gamma_t = \min\left\{C, \frac{l_t}{||\mathbf{x}_t||^2}\right\}, \tag{11}$$

$$l_t = \begin{cases} 0 & \text{if } d_t > 1 \\ 1 - d_t & \text{otherwise} \end{cases}, \tag{12}$$

$$d_t = \sum_{j=0}^{t-1} \Gamma_j \, y_j \, K(\mathbf{x}_j,\mathbf{x}_t). \tag{13}$$

It is interesting to notice that according to this approach, the vectors $\mathbf{x}_t$ could be seen as *support vectors*, as in the Support Vector Machine paradigm (Cristianini and Shawe-Taylor, 2000), because the scope is to maximize the *margin* between relevant and non relevant images. In this approach, only one image per round is randomly drawn from the set of the previous retrieved images, and it can be drawn more than once. t indicates the number of update rounds executed so far, and obviously it is equal to the number of images that have been evaluated until that moment.

4 EXPERIMENTS AND RESULTS

This section contains the datasets description, the evaluation protocol together with the results and comparison of the different algorithms.

4.1 Datasets

Experiments have been carried out using two datasets, namely the Caltech-101, and the Caltech-256 dataset, both from the California Institute of Technology[1]. These datasets are widely used in object recognition, and examples of the images in these datasets can be found by visiting the website. The first dataset consists of 30607 images subdivided into 257 semantic classes, the second one is composed by 9144 images subdivided into 102 semantic classes. Experiments have been carried out by using different descriptors, and results are shown for the *Edge Histogram* descriptor (80 components) (MPE, 2003), and the *Color and Edge Directivity Descriptor* (*Cedd*, 144 components) (Chatzichristofis and Boutalis, 2008). The open source library LIRE (Lucene Image REtrieval) has been used for feature extraction (Lux and Chatzichristofis, 2008).

4.2 Experimental Setup

In order to test the performances of the proposed approaches, 500 query images have been randomly extracted from each dataset, so that they cover all the semantic classes. The top twenty best scored images for each query are returned to the user. Relevance feedback is performed by marking images belonging to the same class of the query as relevant, and all other images in the top twenty as non-relevant. Performances are evaluated in terms of precision, and truncated average precision (Nie et al., 2012). Precision is evaluated by taking into account the top twenty best scored images at each iteration, while the truncated average precision takes into account the ranking of the top T results, i.e., the average precision at depth T:

$$AP@T = \frac{1}{T} \sum_{i=1}^{T} rel(\tau(i)) \frac{\sum_{j=1}^{i} rel(\tau(j))}{i} \tag{14}$$

where $\tau(i)$ is the image at the rank i, and $rel(\tau(i))$ is the associated binary relevance label equal to 1 if $\tau(i)$ is relevant with respect to the query, and 0 otherwise, the highest the value of AP@T, the better the ranking. We performed nine rounds of relevance feedback, so we fixed $T = min(|C_i|, 180)$, where $|C_i|$ is the size of the class of the query, and 180 is the maximum number of images the user may be asked to mark after all the feedback rounds.

In order to choose the most suitable values of the parameters discussed in Sections 2 and 3, a number of preliminary experiments have been performed. Accordingly, we fixed the number of update round t at

[1] http://www.vision.caltech.edu/archive.html

100, we used the RBF kernel in the kernel version of the PA algorithm with a value of σ^2 set to 0.1. We have also tested different values for the parameter C between 0.001 to 1000. The best results have been obtained using $C = 1$. For comparison purposes, relevance feedback has been also computed by a SVM classifier with an RBF kernel, and by the Relevance Score (RS) method, where images are ranked according to the ratio of the distances from the nearest relevant and non-relevant images (Giacinto, 2007).

4.3 Results

Reported results show that the linear formulation of the PA technique allows attaining the highest performance in all the experiments at the end of the feedback rounds. In particular, the precision is quite close to the one attained by RS, while it is greater than the precision attained by the SVM classifier after a few iterations. On the other hand, the analysis of the performances in terms of the AP@T, shows that the linear PA approach allows attaining the highest performances after four iterations, with a significant gap from the performance of the SVM approach. Thus, it can concluded that the linear PA approach allows better exploiting feedback from the user with respect to traditional SVM classification approaches when the amount of available information increases. The linear PA approach also provides higher performances in terms of AP@T than those attained by the RS approach, the difference being smaller than the one between the linear PA approach and the SVM approach.

This behavior can be explained by the similar rationale behind the PA and the RS approaches. Both are aimed at producing a score that allows producing a better ranking of the images, while the SVM approach is aimed at estimating a discriminating surface, without taking into account the relative ranking.

By inspecting the performances attained by the kernel formulation of the PA approach, it can be seen that if just one iteration is allowed, it provides better performances than those of the linear formulation, and in some cases they are the highest. On the other hand, it can be seen that after the second iteration they are poorer than the ones attained by the linear formulation. Thus, the kernel formulation of the PA approach does not provide the same power of the linear formulation in exploiting the feedback information, the reason being the strong relationship with the classification formulation of the problem that turns out not to be the most suited approach for relevance feedback. If the performances of the kernel formulation of the PA approach are compared to the ones provided by SVM, it can be seen that they depend on the feature space employed. In particular the kernel PA approach outperforms SVM when the EH features are considered, while SVM is superior when the CEDD features are used.

5 CONCLUSION

The PA approach can be used to exploit relevance feedback in content based retrieval. In particular, the linear formulation provides good performances, when the user provides a significant amount of feedback information. On the other hand, when few feedback iterations are allowed, the performances are slightly worse than the ones provided by other mechanisms. Anyway, if the user wishes to provide more feedback, the linear PA approach allows improving retrieval performances faster than other mechanisms.

REFERENCES

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

Chatzichristofis, S. A. and Boutalis, Y. S. (2008). Cedd: Color and edge directivity descriptor: A compact descriptor for image indexing and retrieval. In *Lecture Notes in Computer Science*, v. 5008, pp. 312–322. Springer.

Crammer, K., Dekel, O., Keshet, J., Shalev-Shwartz, S., and Singer, Y. (2006). Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Deselaers, T., Paredes, R., Vidal, E., and Ney, H. (2008). Learning weighted distances for relevance feedback in image retrieval. In *ICPR*, pages 1–4. IEEE.

Giacinto, G. (2007). A nearest-neighbor approach to relevance feedback in content based image retrieval. In *CIVR '07*, pp. 456–463, New York, NY, USA. ACM.

Grangier, D. and Bengio, S. (2008). A discriminative kernel-based approach to rank images from text queries. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(8):1371–1384.

Lux, M. and Chatzichristofis, S. A. (2008). Lire: lucene image retrieval: an extensible java cbir library. In *MM '08: Proc.g of the 16th ACM Int. Conf. on Multimedia*, pages 1085–1088, New York, NY, USA. ACM.

Nie, L., Wang, M., Zha, Z.-J., and Chua, T.-S. (2012). Oracle in image search: A content-based approach to performance prediction. *ACM Trans. Inf. Syst.*, 30(2):13.

Paredes, R., Ulges, A., and Breuel, T. (2009). Fast discriminative linear models for scalable video tagging. *Mach. Lear. and Applications, 4th Int. Conf. on*, 0:571–576.

Zhou, X. S. and Huang, T. S. (2003). Relevance feedback in image retrieval: A comprehensive review. *Multimedia Syst.*, 8(6):536–544.

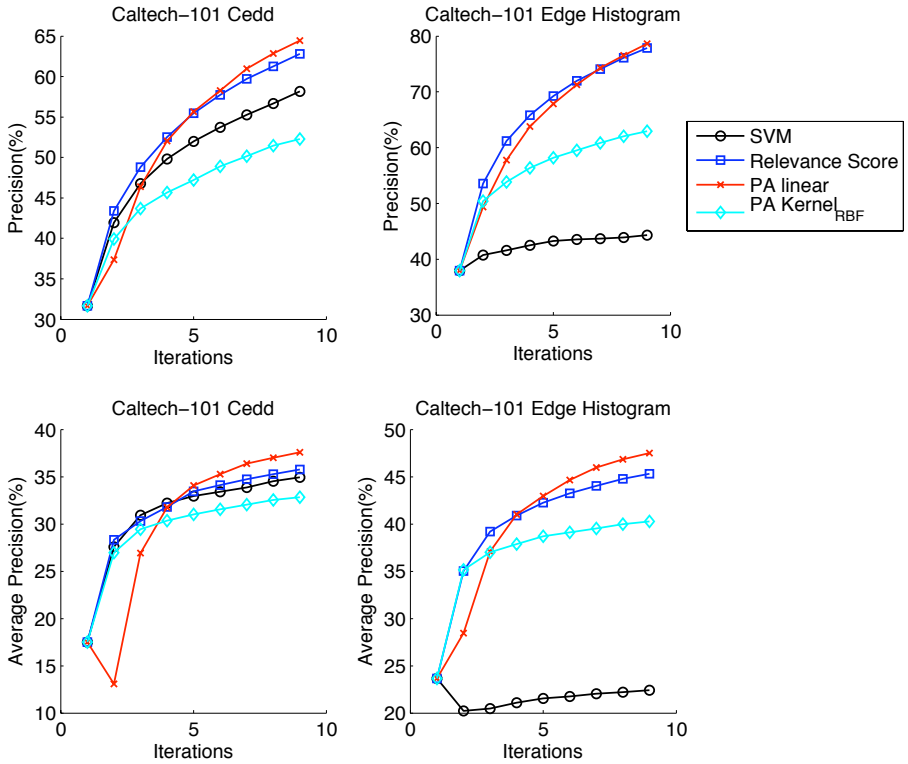

Figure 1: Caltech 101 Dataset - Precision and Average Precision for 9 rounds of relevance feedback using Color and Edge Directivity Descriptor (on the left) and Edge Histogram descriptor (on the right).

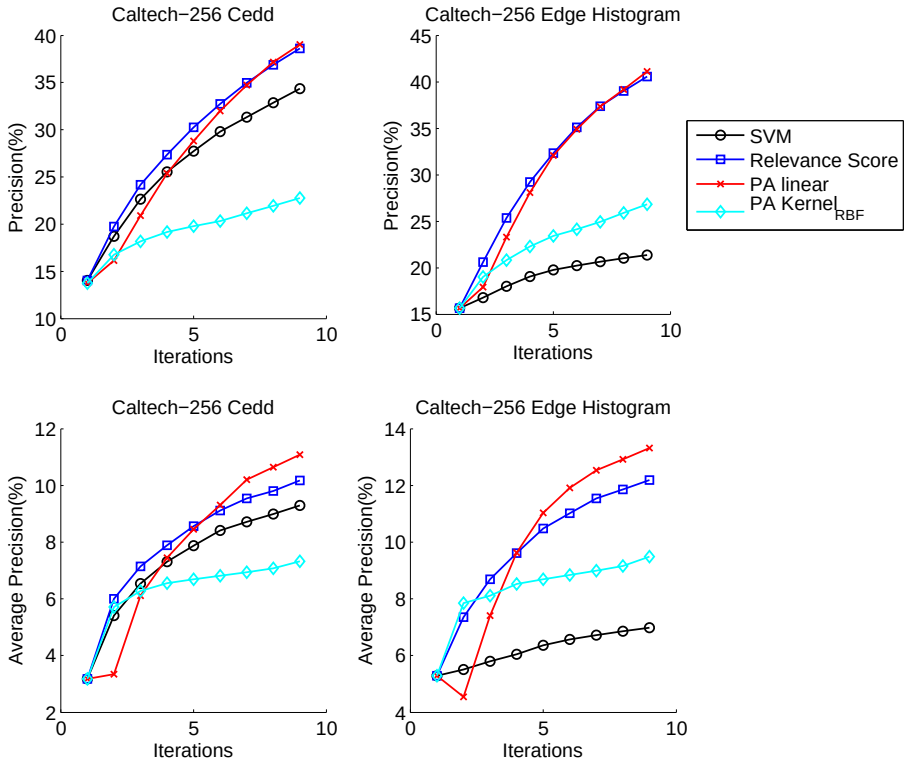

Figure 2: Caltech 256 Dataset - Precision and Average Precision for 9 rounds of relevance feedback using Color and Edge Directivity Descriptor (on the left) and Edge Histogram descriptor (on the right).