# Incorporating Prior Knowledge into SVM for Image Retrieval

Lei Wang, Ping Xue[†], Kap Luk Chan[‡]
*School of Electrical and Electronic Engineering*
*Nanyang Technological University*
*Singapore, 639798*
*E-mail:* $\{elwang, epxue^{\dagger}, eklchan^{\ddagger}\}$@$ntu.edu.sg$

## Abstract

*SVM based image retrieval suffers from the scarcity of labelled samples. In this paper, this problem is solved by incorporating prior knowledge into SVM. Firstly, some prior knowledge of image retrieval is discussed and constructed. After that, the knowledge is incorporated into SVM optimization as a constraint, and a new knowledge-based target function is formulated. Based on this, a framework of image retrieval with knowledge based SVM is proposed. Experimental results demonstrate that the proposed method can effectively improve the learning and retrieval performance of SVM, especially when the number of labelled samples is small.*

## 1. Introduction

Recently, Support Vector Machine (SVM) has been applied to Content-Based Image Retrieval to learn user preference [2, 5, 6, 7, 8]. In this scenario, a user is often asked to provide some "positive" and "negative" image examples via relevance feedback for retrieval. These labelled image samples are then used as training data to train an SVM classifier to perform a bi-class classification over an image database. Those images having the larger positive decision values are retrieved as positive images.

However, the number of labelled samples is often small because a user is unwilling to label too many images in a retrieval. Because learning performance of SVM often degrades with the decreasing number of training samples, poor retrieval performance is resulted. It is especially true for the initial stage of retrieval where the number of training samples is very limited, and the worse is that the retrieval performance at this stage is often a crucial factor on whether the users wish to continue with the search. Some methods have been proposed to get around this problem. In [6], the user is asked to label more images to expand the labelled sample set. However, it increases the burden on users. In [2, 8], Euclidean search is combined with SVM for retrieval. However, the poor learning performance of SVM is left untouched. In [7], transductive SVM is used to improve the learning performance by incorporating unlabelled data. Nevertheless, it is also observed that the performance of this method may be unstable in some cases. Furthermore, whether unlabelled data is truly helpful for supervised learning is still unclear [9].

It has been well realized in the field of machine learning that incorporating prior knowledge helps a learning machine achieving more accurate generalization, especially when training samples are scarce [3]. As a practical application, image retrieval has its own characteristics and prior knowledge. To our best knowledge, no papers have addressed incorporating prior knowledge of image retrieval to boost the retrieval performance of SVM. In this paper, some prior knowledge of image retrieval is discussed and constructed. The knowledge is then incorporated into SVM optimization as a constraint, and a new target function is formulated. Based on this, a framework of image retrieval with knowledge based SVM is proposed. We show, through experimental results, that by incorporating simple prior knowledge, the generalization accuracy of SVM can be well improved, leading to an overall better retrieval performance.

## 2. Support Vector Machine

Let $\mathcal{D} = \{(\mathbf{x}, y)\} \in (\mathbb{R}^n \times \mathcal{Y})^{|\mathcal{D}|}$ be a training set, where $\mathbb{R}^n$ denotes an $n$-dimensional input space, $\mathcal{Y} = \{\pm 1\}$ denotes the label set of $\mathbf{x}$, and $|\mathcal{D}|$ is the size of $\mathcal{D}$. Given $\mathcal{D}$, SVM finds an optimal separating hyperplane, $\langle \mathbf{w}^*, \mathbf{x} \rangle + b^* = 0$, which classifies the two classes by the minimal expected test error. $\mathbf{w}^*$ and $b^*$ are found by minimizing

$$\begin{aligned} & \Phi(\mathbf{w}) = \tfrac{1}{2}\|\mathbf{w}\|^2 + C\textstyle\sum_{i=1}^{|\mathcal{D}|} \xi_i^p \\ subject\ to: \quad & y_i\left(\langle \mathbf{w}, \mathbf{x}_i \rangle + b\right) \geq 1 - \xi_i, \quad i = 1, \cdots, |\mathcal{D}| \end{aligned} \tag{1}$$

Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)
1051-4651/04 $ 20.00 IEEE

where $\xi_i$ ($\xi_i \geq 0$) is the $i$-th slack variable and $C$ is the parameter controlling the trade-off between function complexity and training error. The solution of (1) is $\mathbf{w}^* = \sum_{i=1}^{|\mathcal{D}|} \alpha_i y_i \mathbf{x}_i$ and $b^* = 1 - \langle \mathbf{w}^*, \mathbf{x}^s \rangle$, where $\alpha_i$ is a non-negative real number and $\mathbf{x}^s$ denotes a support vector. In this way, the optimal separating hyperplane is

$$\langle \mathbf{w}^*, \mathbf{x} \rangle + b^* = \sum_{i=1}^{|\mathcal{D}|} \alpha_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle + b^* = 0 \qquad (2)$$

In test phase, $f(\mathbf{x}) = \langle \mathbf{w}^*, \mathbf{x} \rangle + b^*$ is used as decision function, and a test sample is labelled as $\mathrm{sgn}\,[f(\mathbf{x})]$, where $\mathrm{sgn}(\cdot)$ denotes the sign function. The *kernel* trick can be conveniently embedded into SVM to handle more complex patterns.

It can be seen that SVM learns classification rules solely from training samples. When training samples are sufficient, SVM can offer excellent classification performance. However, it may fail if the requirement is not satisfied.

# 3. The proposed method

## 3.1. Prior knowledge in image retrieval

An image retrieval system includes an image database, visual features, query, user, similarity metric, and so on. For each of them, there is some prior knowledge. Since the performance of SVM is mainly related to training data, the prior knowledge on the images in a database is our focus here.

Let $N$ be the number of images in a database, and let $N_p$ and $N_n$ denote the numbers of positives and negatives for a given query, respectively. There is $N = N_p + N_n$ for bi-class classification. For an image, $\mathbf{x}$, its *priori* probability with respect to the negative class can be calculated as $P_{pri}(\mathrm{Neg}|\mathbf{x}) = \frac{N_n}{N}$. Commonly, in an image database, the number of negative images is much larger than the one of the positives. This is especially true for real applications. In this way, it can be obtained that $N_n \gg 0.5N$. As a result, there is

$$P_{pri}(\mathrm{Neg}|\mathbf{x}) \gg 0.5 \qquad (3)$$

This constructs the following prior knowledge that

> $PK_1$: *For a randomly sampled unlabelled image, before seeing any additional evidence, its label is negative with high probability.*

Another prior knowledge can be obtained from long-term learning, a hot topic in recent research on image retrieval. Some algorithms have been devised to predict the labels of the images in a database for a given query by using long-term retrieval history only [1]. Although the predicted labels may not be completely consistent with those in the current user's mind due to the *subjectivity* of human perception, they are still helpful because of the *commonness* of perception. It brings forth the following prior knowledge that

> $PK_2$: *For a randomly sampled unlabelled image, its probabilities of being positive and negative can be predicted, based only on the given query and retrieval history, as* $P_{pri}(\mathrm{Pos}|\mathbf{x})$ *and* $P_{pri}(\mathrm{Neg}|\mathbf{x})$.

In this work, the knowledge $PK_1$ is utilized though the following proposed method can deal with both. We aim to show that the learning and retrieval performance of SVM can be well improved even by incorporating such simple prior knowledge.

## 3.2. Incorporating prior knowledge

The task of incorporating prior knowledge can be modelled as follows:

Given:

- Labelled data set, $\mathcal{D} = \{(\mathbf{x}, y)\} \in (\mathbb{R}^n \times \mathcal{Y})^{|\mathcal{D}|}$;
- Prior knowledge set, $B = \{PK_1, PK_2, \cdots, PK_k\}$;
- A learning machine, SVM.

Find:

- $f(\mathbf{x}) = \langle \mathbf{w}^*, \mathbf{x} \rangle + b^*$, which best fits both $\mathcal{D}$ and $B$.

This can be further expressed as

$$\{\mathbf{w}^*, b^*\} = \arg\min\,[\lambda_{\mathcal{D}} L(f, \mathcal{D}) + \lambda_B L(f, B)] \qquad (4)$$

where $L(f, \mathcal{D})$ denotes the loss resulted from the disagreement between the sign of $f(\mathbf{x})$ ($\mathbf{x} \in \mathcal{D}$) and the label of $\mathbf{x}$ given in $\mathcal{D}$, and $L(f, B)$ denotes the loss due to the disagreement between the sign of $f$ and the labels estimated from $B$. $\lambda_{\mathcal{D}}$ and $\lambda_B$ are their weights, respectively. Both prior knowledge, $PK_1$ and $PK_2$, can be used to assign *pseudo*-labels to unlabelled images, and a *pseudo*-labelled data set, $\mathcal{D}_B = \{(\mathbf{x}, y_B)\}$, can be obtained. In this way, equation (4) becomes

$$\{\mathbf{w}^*, b^*\} = \arg\min\,[\lambda_{\mathcal{D}} L(f, \mathcal{D}) + \lambda_B L(f, \mathcal{D}_B)] \qquad (5)$$

This means to train $f$ on both $\mathcal{D}$ and $\mathcal{D}_B$ while different penalties for training error are assigned. Based on equation (1), the target function for SVM incorporating prior knowledge can be proposed as

$$\begin{aligned} &\Phi_B(\mathbf{w}) = \tfrac{1}{2}\|\mathbf{w}\|^2 + C_{\mathcal{D}} \textstyle\sum_{i=1}^{|\mathcal{D}|} \xi_{\mathcal{D},i}^p + \sum_{j=1}^{|B|} C_{B,j} \xi_{B,j}^p \\ &subject\ to: \\ &y_{\mathcal{D},i}\,(\langle \mathbf{w}, \mathbf{x}_{\mathcal{D},i} \rangle + b) \geq 1 - \xi_{\mathcal{D},i}, \quad i = 1, \cdots, |\mathcal{D}| \\ &y_{B,j}\,(\langle \mathbf{w}, \mathbf{x}_{B,j} \rangle + b) \geq 1 - \xi_{B,j}, \quad j = 1, \cdots, |\mathcal{D}_B| \end{aligned} \qquad (6)$$

where $C_{\mathcal{D}}$ and $C_{B,j}$ ($j = 1, \cdots, |\mathcal{D}_B|$) are the parameters for $\mathcal{D}$ and $\mathcal{D}_B$, respectively. $C_{B,j}$ varies with the confidence of the *pseudo*-label for $\mathbf{x}_{B,j}$. When the knowledge like

$PK_1$ is applied, $C_{B,j}$ is identical to $\mathbf{x}_{B,j}, j = 1, \cdots, |\mathcal{D}_B|$. For knowledge as $PK_2$, $C_{B,j}$ will be different for different data. Also, a $C_{\mathcal{D},i}$ can be applied if the confidence information of $y_{\mathcal{D},i}$ is available. Here, it is simply considered as $C_{\mathcal{D}}$. A large $C$ means a big penalty for training error. The relationship between $C_{\mathcal{D}}$ and $C_B$ controls the trade-off between the training errors on $\mathcal{D}$ and $\mathcal{D}_B$. It can be set by a retrieval system according to user preference. The optimization in equation (6) can be easily solved by the tool packages for the original SVM.

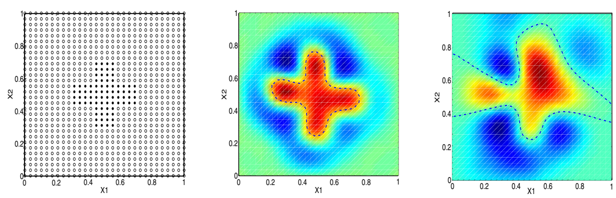
By incorporating prior knowledge, SVM can effectively avoid being misled by insufficient training data when it generalizes over them. It can be expected to achieve better learning and retrieval performance when training samples are insufficient. One potential disadvantage is an increase on training time due to incorporating *pseudo*-labelled data. In the current work, this problem is tackled by randomly sampling a part of *pseudo*-labelled data for training.

A framework of image retrieval with the knowledge based SVM is proposed as follows. (1)A user launches a retrieval with query examples. These examples form a labelled data set, $\mathcal{D}$; (2)According to prior knowledge $B$, a *pseudo*-labelled data set, $\mathcal{D}_B$, is set up; (3)Train SVM on both $\mathcal{D}$ and $\mathcal{D}_B$, and calculate the decision values for the images in the database; (4)If the user will perform more feedbacks, show the selected images to the user to get their labels. Otherwise, sort the images in the database according to the descending order of decision values, and show the top images as the final retrieval result; (5)After the user labels the shown images, add them to $\mathcal{D}$ and remove them from $\mathcal{D}_B$. Then go to step(3).

## 4. Experimental Result

One artificial and one real image databases are used in the experiments. The artificial database is shown in Figure 1(a). The positive "image" class is represented by a cross-shaped cloud of "$*$" while the negative is denoted by "$\circ$". The numbers of positive and negative samples are 80 and 820, respectively. The real image database includes 600 general color images composed from *Corel* Stock Photos. Six image classes are defined based on high-level semantics, and each class includes 100 image samples. Based on the $CIE - Lab$ color space, a feature vector of color moments [4] is defined for each image. Classification accuracy and *Precision* are used to measure the learning and retrieval performance, respectively. In the experiments, Gaussian RBF kernel is used, and the Gaussian width is set as the average Euclidean distance between the nearest positive and negative training samples. The prior knowledge $PK_1$ is used, and all the unlabelled images are labelled as *negative*. The parameters $C_{\mathcal{D}}$ and $C_{B,j}$ are experimentally set to $10^3$ and $10^0$, respectively. Note that this setting strikes a balance between incorporating prior knowledge and preventing the inaccurate inference of the knowledge (for example, some unlabelled true positive images may be wrongly labelled as negative under $PK_1$) from overriding the information from user-labelled data.

Figure 2(a), (b), and (c) show the comparison between the proposed method, SVM+PK, and SVM on the artificial database when different numbers of labelled data are available. All the results are averaged over 100 groups of experiments. From the sub-figure(a), it can be seen that SVM+PK achieves higher classification accuracy while smaller standard deviation than SVM, and that the less the number of labelled data, the more significant the improvement. This result indicates the benefit of incorporating retrieval prior knowledge. Figure 1(b) and (c) illustrate two classification examples by using SVM+PK and SVM, respectively. The dotted lines highlight the decision boundaries. It can be seen, by referring to Figure 1(a), that the boundary from SVM+PK is more accurate. Figure 2(b) and (c) present the comparison on *Precision* of top 20 and 50 retrieved images, respectively. Compared with SVM, SVM+PK consistently shows better retrieval performance. To reach a given *Precision* level, SVM+PK requires less number of training data than SVM.



(a) (b) (c)

**Figure 1. The artificial database**

Besides one-shot retrieval, SVM+PK and SVM are also compared on relevance feedback based retrieval. The active learning mode in [6] is used. To simulate real cases, only one positive and one negative examples are assumed to be available at the beginning of retrieval, and 20 images are shown in each feedback. Also, this experiment tests the performance of SVM+PK when different percentages of *pseudo*-labelled samples are randomly sampled for training. Figure 2(d), (e), and (f) show the comparison results. Again, SVM+PK shows better classification and retrieval performance (For the case of $1\%$, SVM+PK is comparable to SVM). The more the incorporated *pseudo*-labelled samples, the higher the improvement. In real applications, short response time can be achieved by incorporating less number of *pseudo*-labelled samples. Figure 3(a) and (b) present the comparison on classification accuracy and *Precision* of top 20 over the real database. Similar conclusions are drawn.

For *Precision* of top 10 and 50 retrieved images, SVM+PK still shows better performance. They are not listed due to the limitation of space. Besides these, preliminary experiment is performed to compare SVM+PK with the methods of SVM with Euclidean search [8] (SVM+EU) and Transductive SVM [7] (SVM+TRANS). For SVM+PK and SVM+TRANS, $10\%$ *pseudo*-labelled samples and the same percentage of unlabelled samples are used, respectively. As shown in Figure 3(c) and (d), SVM+PK still shows the best performance.

## 5. Conclusion

In this paper, retrieval prior knowledge is incorporated into SVM to improve its retrieval performance on the limited number of labelled samples. To achieve this, a *pseudo*-labelled data set is set up according to prior knowledge, and SVM is trained on this set and the given labelled data together. The experimental results show that, by using the proposed method, SVM based image retrieval achieves better performance, especially when the number of labelled samples is small.

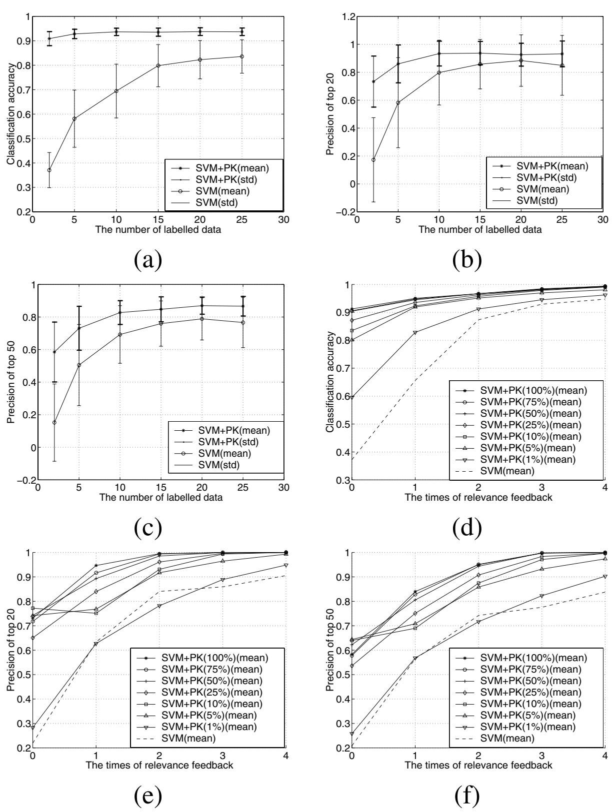



**Figure 2. Comparison on artificial database**

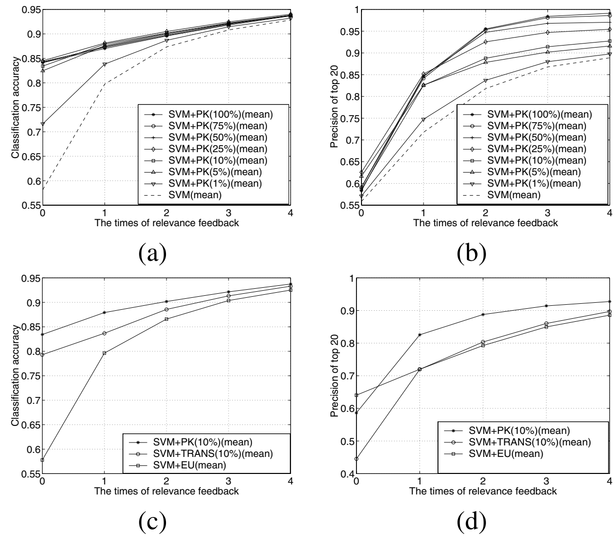



**Figure 3. Comparison on real database**

## References


[1] A. Dong and B. Bhanu. A New Semi-supervised EM Algorithm for Image Retrieval. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[2] G. Guo, A. K. Jain, W.-Y. Ma, and H. Zhang. Learning Similarity Measure for Natural Image Retrieval with Relevance Feedback. *IEEE Transactions on Neural Networks*, 13(4):811–820, 2002.

[3] T. M. Mitchell. *Machine Learning*. McGraw-Hill Companies, Inc, 1997.

[4] M. Stricker and M. Orengo. Similarity of Color Images. *Proceedings of SPIE Storage and Retrieval for Image and Video Databases*, 2420:381–392, 1995.

[5] Q. Tian, P. Hong, and T. S. Huang. Update Relevant Image Weights for Content-Based Image Retrieval using Support Vector Machines. *IEEE International Conference on Multimedia and Expo (II)*, pages 1199–1202, 2000.

[6] S. Tong and E. Chang. Support Vector Machine Active Learning for Image Retrieval. *Proceedings of ACM International Conference on Multimedia*, pages 107–118, Ottawa, Canada, Oct, 2001.

[7] L. Wang and K. L. Chan. Bootstrapping SVM Active Learning by Incorporating Unlabelled Images for Image Retrieval. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2003.

[8] L. Wang, K. L. Chan, and Y. P. Tan. Image Retrieval with SVM Active Learning Embedding Euclidean Search. *Proceedings of IEEE International Conference on Image Processing (ICIP)*, 2003.

[9] T. Zhang and F. Oles. A Probability Analysis on The Value of Unlabeled Data for Classification Problems. *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pages 1191–1198, 2000.