Knowing a Tree from the Forest: Art Image Retrieval using a Society of Profiles

Kai Yu[†], Wei-Ying Ma[‡], Volker Tresp[§], Zhao Xu[†], Xiaofei He[£],
HongJiang Zhang[‡], Hans-Peter Kriegel[†]
[†]Institute for Computer Science, University of Munich, Germany
[‡]Microsoft Research Asia, Beijing, China
[§]Corporate Technology, Siemens AG, Munich, Germany
[£]Department of Computer Science, University of Chicago, USA

ABSTRACT

This paper aims to address the problem of art image retrieval (AIR), which aims to help users find their favorite painting images. AIR is of great interests to us because of its application potentials and interesting research challenges—the retrieval is not only based on painting contents or styles, but also heavily based on user *preference profiles*. This paper describes the collaborative ensemble learning, a novel statistical learning approach to this task. It at first applies probabilistic support vector machines (SVMs) to model each individual user's profile based on given examples, i.e. liked or disliked paintings. Due to the high complexity of profile modelling, the SVMs can be rather *weak* in predicting preferences for new paintings. To overcome this problem, we combine a society of users' profiles, represented by their respective SVM models, to predict a given user's preferences for painting images. We demonstrate that the combination scheme is embedded in a Bayesian framework and retains intuitive interpretations—like-minded users are likely to share similar preferences. We report extensive empirical studies based on two experimental settings. The first one includes some controlled simulations performed on 4533 painting images. In the second setting, we report evaluations based on user preferences collected through an online web-based survey. Both experiments demonstrate that the proposed approach achieves excellent performance in terms of capturing a user's diverse preferences.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*information filtering, retrieval models*

General Terms

Algorithms, Theory, Human Factors

MM'03, November 2–8, 2003, Berkeley, California, USA.

1. INTRODUCTION

The explosive growth of digital media and its usage on the web has opened both great challenges and opportunities to computer science researchers. One such example is the lively research area content-based image retrieval (CBIR) [7, 9, 17, 25, 6], which aims to effectively and efficiently locate a few relevant images in a large database according to a user's *query concept*. This paper will focus on another related but different topic, art image retrieval (AIR), which is of great interests to us because of its application potentials and technical challenges. In recent years many museums, galleries or commercial companies have put their art images on the web for on-line exhibitions or selling. The WWW technology has created dramatic flexibilities in helping people freely share, exchange, enjoy and purchase paintings. For example, POSTERSHOP.com is an on-line poster seller, who presents about 22,000 art images on its web site, ranging from classical to modern paintings. Based upon demands from users, the company can send high-quality printouts to them. To facilitate the search of paintings, POSTSHOP.com categorized all the art images by styles (e.g. classical art, impressionism, ...) or other information like name of artists. However, the perception of art works are highly governed by users' personalities. Category-based search might not be sufficient in capturing users' personal interests. In this paper we will propose an effective machine learning approach to helping people find favorite art images from a large art image collection.

1.1 Properties of Art Image Retrieval

From technical point of view, AIR encounters many challenges that are also common in CBIR, like high-dimension feature space, nonlinear distributions, insufficient training examples and the gap between low-level features and high-level contents. Thus previous research on CBIR has provided a good basis for our work on AIR. Moreover, there are some other unique characteristics coming with AIR. In the following we will discuss some major issues.

- *A Profile-Driven Process*: One major change from CBIR to AIR is that, the *query concept* is replaced by the *user profile*[1], which is governing the entire retrieval process. For example, in CBIR a user may raise a query "find all images containing flowers", while in

[1]In this paper the term "profile" is used to generally describe a user's preference for art images.

Figure 1: Starry Night (Vincent van Gogh, 1889) (1) Low-level visual features: blue, dark, yellow, strong trace of hand drawing; (2) High-level semantic content: sky, stars, tree, houses, mountain; (3) Higher-level abstract properties: adventurous use of color and brush, tumultuous night sky, etc.

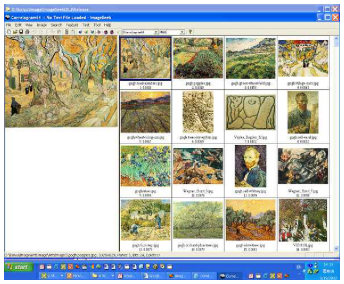

Figure 2: We use a Vincent van Gogh's painting as the example to find similar images. The result is surprising: 11 Gogh's paintings are found among top 15 returned images. The distance measure is Euclidian distance based on feature *correlagram64*, which carries both color and texture information.

AIR a user may just implicitly keep a query in mind "return some nice images to decorate my house". In the first case "flowers" is a query concept that conveys explicit information of contents. The search criterion (in this example, "flowers") is somewhat "hard" in the sense that all the people agree what kind of images should be returned. While in the AIR case the query has no hard requirement for contents but highly depends on the profile of the user. We notice that AIR is more profile-driven and the search criterion can not be held by all be people but possibly by only *a small community of like-minded users*. This notice will serve as one important basis for this paper and greatly inspires the proposed algorithm.

- *A New Gap*: As indicated in Fig. 1, there is another gap between the low-level visual features (e.g. color and texture) and the *higher-level abstract properties*, e.g. painting styles and expressed feelings or emotions. In some sense, this gap might be more important for AIR since the higher-level abstract properties are always more indicative for the expression of art images (especially for some modern art images).

- *Diversity of User Preferences*: As the example shown in Fig. 2, to some extent, low-level features like color and texture reflect some important characteristics of paintings. However, as shown in Fig. 3, a user's interests in art images is typically diverse. One may love different styles of paintings in a meanwhile. This paper assumes that a user's preference for art images can be described as a union of disjoint regions in the low-level feature space, as shown in Fig. 3, where consistency of interests is only held in a local region. Under this assumption, this paper is facing a challenge—given a small set of examples that only partially conveys a user's interests, how can we infer the user's overall preferences?

- *Uncertainty of User Preferences*: Again, a user will normally give a very hard relevance judgment for an image given a query concept. However, people are always not so confident with their preferences for art images. In our experiments we had the following observation: We re-present art images to some users who

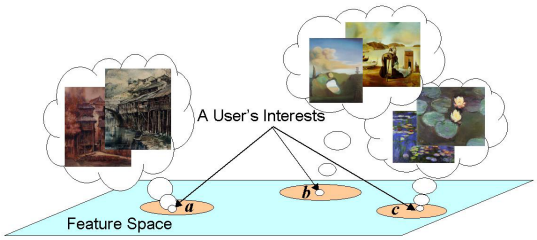

Figure 3: A user's diverse preferences for art images, represented as union of disjoint regions in visual feature space, i.e. a nonlinear distribution. A question is, if a user gives training examples only distributed in region *a*, how to infer the user's preferences for region *b* or *c*?

marked the same images two days ago, then we observed that many of them changed their opinions for some of images. The case indicates that we should find effective means, e.g. statistics, to model the uncertainty of user preferences.

1.2 Two Possible Approaches to AIR

We can adopt the general idea of CBIR, which trains a model based on a limited number of examples (i.e. liked and disliked images) provided by a user, and then use the trained model to predict the user's preferences for other unseen images.[2] The learning models can be either SVM[29], discriminant analysis[35], boosting[28], or similarity search[15]. However, we may often encounter three problems when applying CBIR methods to AIR: (1) CBIR methods are mainly built upon low-level visual features (e.g. color, texture, and shape etc.), which are poor indicators of painting images; (2) Derived models based on small training set and weak visual features are easily over-fitting and often make poor predictions; (3) Due to the diversity of user interests, images fitting in an active user's interests but different from given examples in feature space can not be found (As illustrated in Fig. 3). In information retrieval literature, this family of methods are often referred as *content-based filtering*.

[2]In the following parts, we call the user to be predicted as "active user".

The other choice is *collaborative filtering*(CF) [24, 26], which uses like-minded users' opinions to predict the active user's preferences. This method seems to be suitable to our task, because, as we pointed out in Sec. 1.1, AIR is more profile-driven and the search criterion is possibly held in a small community of like-minded users. CF ignores the content of items (i.e. art images in this paper) and thus avoids the difficulty of low-level features and can potentially find diverse results by taking other users' advices. However, It greatly suffers from the "new item problem"—It can not make predictions for images on which nobody has expressed opinions at all. For a large and ever-growing image database, user annotations are typically very sparse and thus plenty of images can be "new items".

1.3 Overview of the Presented Approach

In this paper we will describe an algorithm, called *collaborative ensemble learning*, to dramatically overcome the weaknesses of existing algorithms and provide a principled solution to the AIR task. It firstly applies a probabilistic SVM (PSVM) to model each user's profile, based on a few examples of liked and disliked painting images given by the user. The probabilistic formalism of SVM with RBF (radius basis function) kernels allows us to naturally handle the nonlinearity of distributions and the uncertainty of user preferences. More importantly, unlike conventional CBIR methods, collaborative ensemble learning further combines a "forest" of previously learned other users' profiles (i.e. PSVM models) to build a mixture model, which acts as some kind of PSVM ensemble to predict the active user's preferences. The ideas of collaborative ensemble learning can be explained as the following.

- Collaborative ensemble learning uses probabilities to encode the *like-mindedness* between users, and derives a weighted combination of profiling models to make predictions for active users. The idea is similar with collaborative filtering (CF), but new images can be handled now. Collaborative ensemble learning is a unified solution combining CF and CBIR. The approach is naturally derived from a Bayesian framework, which demonstrates the theoretical soundness of the solution.

- To some extent, it merges the gap between low-level features and higher-level abstract properties. Since each user's preference for an image actually conveys some higher-level properties of the image, combining a large community of profiles can help AIR systems effectively understand art images from human perceptual perspectives.

- It can capture a user's diverse interests based on a small training set. Learning the interests of users typically requires many training examples. Unfortunately users are typically impatient to give examples. Incorporation of other advisory profiles through a weighted combination scheme can be viewed as a way to "augment" the training data.

1.4 Organization of This Paper

The rest of this paper is organized as follows. We will describe the idea of modelling user profiles with probabilistic SVMs in Sec. 2. Then in Sec. 3 we first present a fully Bayesian framework to art image retrieval (AIR) and then derive the proposed collaborative ensemble learning based on PSVMs to predict user preferences for art images. In Sec. 4 we report the empirical evaluation of the algorithm on two sets of art image data. After a brief introduction to related work in Sec. 5, we end by giving conclusions and an outlook to future work in Sec. 6.

2. MODELLING USER PROFILES WITH PROBABILISTIC SVMS

In this section we will present probabilistic SVMs (PSVM) to model user profiles, i.e. user preferences for art images. PSVM has some appealing features, while it suffers the problem of high variances caused by high diversity of user preferences and insufficient training data. This explains why we propose major extensions in Sec. 3.

2.1 Notation

Suppose a database of art images has been given, where each of the art images (out of a total of M) is represented as a vector of features $\boldsymbol{x}_j$ $j = 1, \ldots, M$. Similarly, we have collected preference data for different users, where the preference data is a set of rated art images together with an opinion $+1$ (liked that particular image) or -1 (disliked). We consider a total of L users. Each user i has given ratings for a set of art images, denoted by $\mathcal{R}_i$, where art image $j \in \mathcal{R}_i$ has been rated with value $y_{i,j}$. Thus the examples given by user i, $i = 1, \ldots, L$, is denoted by the set $\mathcal{D}_i = \{(\boldsymbol{x}_j, y_{i,j}) | j \in \mathcal{R}_i,\ y_{i,j} \in \{+1, -1\}\}$. In general, assuming we can model user i's profile by a parametric model θ_i, then user i's preference for art image $\boldsymbol{x}$ can be given by $p(y = +1|\boldsymbol{x}, \theta_i)$, which indicates the probability of that user i likes art images $\boldsymbol{x}$. In this paper, we will use probabilistic SVMs to form the profile model θ_i.

2.2 Support Vector Machines

Support vector machines (SVMs) are a classification technique with strong backing in statistical learning theory [30]. It has been applied with great success in many challenging classification problems, including text categorization [16] and image retrieval [29].

We consider SVMs for learning the preferences of one particular user i, based on the examples $\mathcal{D}_i$ this user has previously provided. A standard SVM would predict user i's preferences on some art images $\boldsymbol{x}$, represented by its feature vector, by forming a weighted combination as follows

$$y = \text{sign}(f^i(\boldsymbol{x})) = \text{sign}\Big(\sum_{j \in \mathcal{R}_i} y_{i,j}\alpha_{i,j}k(\boldsymbol{x}_j, \boldsymbol{x}) + b_i\Big) \qquad (1)$$

We will later use θ_i to stand for the SVM profile model for user i, with θ_i containing all SVM model parameters $\alpha_{i,j}$ and b_i.

During training, the weight parameters $\alpha_{i,j}$ of the SVM are determined by minimizing the cost function

$$C \sum_{j \in \mathcal{R}_i} (1 - y_{i,j} f^i(\boldsymbol{x}_j))_+ + \frac{1}{2}\boldsymbol{\alpha}_i^T K^i \boldsymbol{\alpha}_i \qquad (2)$$

By $(\cdot)_+$, we denote a function with $(x)_+ = x$ for positive x, and $(x)_+ = 0$ otherwise. K^i is the matrix of all pairwise kernel evaluations on the training data $\mathcal{D}_i$, and $\boldsymbol{\alpha}_i$ is a vector containing all parameters $\alpha_{i,j}$.

2.3 Probabilistic Extensions to SVMs

In their standard formulation, SVMs only output a prediction $+1$ or -1, without any associated measure of confidence. This paper will consider a special modification of SVM, which can output *a posteriori* class probabilities. This modification retains the powerful generalization ability of SVMs and paves the way to wide extensions, which will be described in the next section. Probabilistic extensions of the SVM, where an associated probability of class membership is output, have been independently suggested by several authors. For our work, we use a probabilistic version of the SVM (PSVM) similar to the one proposed by [21]. Here, the probability of membership in class y, $y \in \{+1, -1\}$ is given by

$$p(y|\boldsymbol{x}, \theta_i) = \frac{1}{1 + \exp(yA_i f^i(\boldsymbol{x}))} \tag{3}$$

A_i is the parameter[3] to determine the slope of the sigmoid function. This modified SVM retains exactly the same decision boundary $f^i(\boldsymbol{x}) = 0$ as defined in Eq. (1), yet allows an easy computation of posterior class probabilities. We use a cross validation scheme to set this parameter A_i for each model. Details and the methodology to select the other parameters for the SVM model will be given along with the appendix section at the end of this paper.

So far we have described a model for the preferences of an individual user, based on probabilistic SVMs. Given some training data containing art images the user likes and dislikes, this model can predict—based on the features of art images—an individual user's preferences. SVM models are known for their excellent performance in many challenging classification problems. However, due to the high complexity of AIR and small training data, the trained models for individual users may have very high variance and only a poor generalization ability. In the following section, we will present a way of combining the individual user models, thus exploiting the knowledge we have gained from possibly like-minded users, to greatly improve the performance of an art image retrieval system.

3. COLLABORATIVE ENSEMBLE LEARNING FOR ART IMAGE RETRIEVAL

In this section we will make major extensions to the profile modelling approach suggested in Sec. 2. The proposed approach will take a strategy of *knowing a tree from the forest*, which combines a society of people's profiles, represented by their respective PSVM models, to make predictions for a query user. The proposed combination scheme is derived from a general Bayesian image retrieval model and retains intuitive explanations.

We particular consider a stage where L users have visited the AIR system and assume that we have collected a set of liked and disliked images for each user i, denoted by $\mathcal{D}_i$, $i = 1, \ldots, L$. For each user, a PSVM model has been built according to Eq. (1). We summarize the parameters for this model by θ_i. We will use the index q to indicate the query user (i.e. active user), and the corresponding training example set $\mathcal{D}_q$.

[3]Platt's original formulation used an additional bias term in the denominator $1 + \exp(yA_i f^i(\boldsymbol{x} + b_i))$. Since we typically only have very few training data available, we restrict the model to containing only one additional parameter.

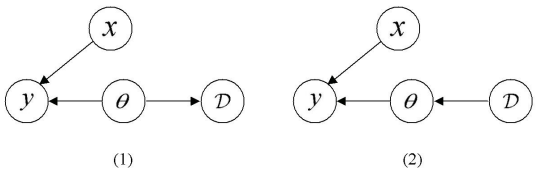

Figure 4: Two equivalent graphical models of art image retrieval

3.1 A Bayesian Model

In this section we will first investigate the problem from a general Bayesian perspective. Given a user's training data $\mathcal{D}_q$, i.e. examples of liked or disliked art images, the final goal of AIR is to predict user q's preference for art image $\boldsymbol{x}$ by:

$$\hat{y} = \underset{y \in \{-1,+1\}}{\arg\max}\ p(y|\boldsymbol{x}, \mathcal{D}_q) \tag{4}$$

To put the retrieval process into a fully probabilistic framework, we make the following assumptions:

1. User profile θ is generated from a prior distribution $p(\theta)$. This prior generally describes the social tastes for art images.
2. Distribution of art images $\boldsymbol{x}$ is independent of user profile θ.
3. User gives examples $\mathcal{D}_q$ based on his or her own profile θ.
4. Given a user's profile, the preferences for art images $\boldsymbol{x}$ are mutually independent.

Then we can use a joint distribution to model the probability of an event that a user q with profile θ who gives examples $\mathcal{D}_q$ has opinions y for an art image $\boldsymbol{x}$:

$$p(\theta, \mathcal{D}_q, \boldsymbol{x}, y) = p(\theta)p(\mathcal{D}_q|\theta)p(y|\boldsymbol{x}, \theta)p(\boldsymbol{x}) \tag{5}$$

It is worth noticing that an equivalent symmetric version of above equation can be obtained as

$$p(\theta, \mathcal{D}_q, \boldsymbol{x}, y) = p(\mathcal{D}_q)p(\theta|\mathcal{D}_q)p(y|\boldsymbol{x}, \theta)p(\boldsymbol{x}) \tag{6}$$

Interestingly, the equivalent version Eq. 6 is parallel with our understanding of retrieval process: Observing the training data $\mathcal{D}_q$ with piror distribution $p(\mathcal{D}_q)$, the system infers the *a posteriori* distribution of user profile and then predicts user's opinions y for any art image $\boldsymbol{x}$ generated from $p(\boldsymbol{x})$. The graphical representations of two equivalent generative models are illustrated in Fig. 4. Through applying Bayes law and integration over $p(\theta)$, we get the following expression:

$$\begin{aligned} p(y|\boldsymbol{x}, \mathcal{D}_q) &= \int p(\theta|\mathcal{D}_q)p(y|\boldsymbol{x}, \theta)d\theta \\ &= \int \frac{p(\mathcal{D}_q|\theta)p(\theta)}{p(\mathcal{D}_q)} p(y|\boldsymbol{x}, \theta)d\theta \\ &= E_\theta\left[\frac{p(\mathcal{D}_q|\theta)}{p(\mathcal{D}_q)} p(y|\boldsymbol{x}, \theta)\right] \end{aligned} \tag{7}$$

where

$$p(\mathcal{D}_q) = \int p(\mathcal{D}_q|\theta)p(\theta)d\theta = E_\theta[p(\mathcal{D}_q|\theta)] \tag{8}$$

In above equations $E_\theta[\cdot]$ denotes the expectation over $p(\theta)$, the prior distribution of user profile θ. Eq. (7) provides a principled way to calculate the conditional probability of user preference y for art image $\boldsymbol{x}$ given observed examples $\mathcal{D}_q$. With combination of Eq. (4), Eq. (7) and Eq. (8), the art image retrieval can be solved in a fully Bayesian framework.

However, fully Bayesian solution is intractable due to two reasons: (1) To calculate the *a posteriori* probability $p(\theta|\mathcal{D}_q)$, we need to know the prior $p(\theta)$, which is hard to assume; (2) Even we can calculate $p(\theta|\mathcal{D}_q)$, the integration Eq. 7 will still be, in general, analytically intractable.

In the following section we will show how a suitable approximation to the prior distribution can be formulated, thereby solving the above two problems.

3.2 Collaborative Ensemble Learning

In this section, we describe a novel method to overcome the problems associated with Eq. (7). Using a simple non-parametric for the prior distribution $p(\theta)$, we find approximate expressions for Eq. (7). The result for predicting a query user's preferences can be interpreted as a combination of a society of user profiles, where like-minded users are given a higher weight. We call this approach *collaborative ensemble learning*.

Collaborative ensemble learning can be introduced in a straight-forward way as follows. We assume a given set of profiles of individual users, $\{\theta_1, \ldots, \theta_L\}$. In general, the user profiles can be modelled in an arbitrary form. In this paper, we restrict ourselves to profiles based on probabilistic SVM models, as introduced in Sec. 2. The goal in the prediction stage is to infer the rating of some active user q on an art image $\boldsymbol{x}$.

We start by noting that the profile θ_i of *every user* can be viewed as an individual sample generated from the same prior distribution $p(\theta)$. $p(\theta)$ should reflect the distribution of actual user profiles. This repeated sampling from $p(\theta)$ allows us to learn a complex prior distribution $p(\theta)$ from the data. More formally, we assume that $p(\theta)$ can be approximated by the empirical distribution[4] of given user profiles in the data base

$$p(\theta) \approx \frac{1}{L} \sum_{i=1}^{L} \delta(\theta - \theta_i)$$

with $\delta(\cdot)$ denoting the Dirac delta function. Given this prior,

$$p(\theta, \mathcal{D}_q, y|\boldsymbol{x}) \approx \frac{1}{L} \sum_{i=1}^{L} \delta(\theta - \theta_i) p(\mathcal{D}_q|\theta) p(y|\theta, \boldsymbol{x})$$

from which follows that

$$p(\mathcal{D}_q, y|\boldsymbol{x}) \approx \frac{1}{L} \sum_{i=1}^{L} p(\mathcal{D}_q|\theta_i) p(y|\theta_i, \boldsymbol{x})$$

such that we obtain for the predicted rating

$$p(y|\boldsymbol{x}, \mathcal{D}_q) \approx \frac{\sum_{i=1}^{L} p(y|\boldsymbol{x}, \theta_i) p(\mathcal{D}_q|\theta_i)}{\sum_{k=1}^{L} p(\mathcal{D}_q|\theta_k)} \tag{9}$$

Since we assume that choosing the example images in $\mathcal{D}_i$ is independent of the user profile, we get for the likelihood terms

$$p(\mathcal{D}_q|\theta_i) = \prod_{j \in \mathcal{R}_q} p(y_{q,j}|\boldsymbol{x}_j, \theta_i) p(\boldsymbol{x}_j) \tag{10}$$

Accordingly, Eq. 9 can be rewritten as follows

$$p(y|\boldsymbol{x}, \mathcal{D}_q) \approx \sum_{i=1}^{L} w_i \cdot p(y|\boldsymbol{x}, \theta_i) \tag{11}$$

where

$$w_i = \frac{\prod_{j \in \mathcal{R}_q} p(y_{q,j}|\boldsymbol{x}_j, \theta_i)}{\sum_{k=1}^{L} \prod_{j \in \mathcal{R}_q} p(y_{q,j}|\boldsymbol{x}_j, \theta_k)} \tag{12}$$

Note that the predicted rating Eq. (7) can be evaluated easily for any kind of models θ for individual user preferences. In this paper, we will use the probabilistic SVM given in Eq. (3) as the model for an individual user's preferences.

3.3 Discussions

An alternative derivation of Eq. (9) can be roughly made by reference to Monte Carlo sampling [8]. Assuming that the accumulated profile models θ_i represent randomly drawn and independent samples from the prior distribution $p(\theta)$, one can directly apply Monte Carlo integration to Eq. (7). The result is exactly the same as given in Eq. (9). Note that restricting hidden variables (in our case, θ) to a set of finite states has been widely adopted in Bayesian inference, i.e. the problem of integrating over the space of hidden variables, for example, constrained Gaussian mixture models [13] and generative topographic mapping [3]. However, a more elegant derivation can be made from a hierarchical Bayesian perspective [32].

Eq. (11) can be interpreted as predictions based on a mixture model with L components. The predicted rating of some given art image $\boldsymbol{x}$ under user i's model, $p(y|\theta_i, \boldsymbol{x})$ takes on the role of a mixture component. The term w_i is the according weight of the component. As indicated in Eq. (12), a higher likelihood of the query user's example data $\mathcal{D}_q$ under some other user i's model indicates that these two persons share similar opinions. Therefore, collaborative ensemble learning simulates the intuition that like-minded people share similar preferences for art images.

One alternative implementation of Eq. (11) is to pick up the K models with largest weights and then make predictions by averaging them, where K is a tuning parameter which can be empirically decided. Since Eq. (12) is essentially proportional to multiply of many likelihood terms, the calculated weights could be improperly scaled if those likelihoods are not precisely estimated. Average of top K models can remove this sensitivity and always lead to stable results in practice. In our experiments, we adopt this way with $K = 20$.

The computational cost of training collaborative ensemble learning is essentially the same as normal content-based image retrieval methods. For each user i, it trains a PSVM model θ_i given the according examples $\mathcal{D}_i$ and then computes the likelihood $p(y|\boldsymbol{x}_j, \theta_i)$ for all the art images $\boldsymbol{x}_j$ in database. If we memorize the computed likelihoods for any i and j, then Eq. (11) and Eq. (12) can be directly calculated in prediction phase. However, this solution may lead to high memory cost. We will discuss this issue in our future work in Sec. 6.

[4]This can also be seen as a Parzen density estimate, with window width going to zero. One might also use window functions with non-vanishing width here, given they allow a simple analytic treatment.

4. EMPIRICAL STUDY

Empirical evaluations of our learning method are conducted in the following two experimental settings:

- *Simulation on 4533 painting images.* From *Meisterwerke der Malerei* CDs we collected 4533 painting images, covering antique Egyptian and Arab frescos, Chinese traditional paintings, India arts, European classical paintings, impressionism paintings, and modern arts in early 1900s. To enable an extensive objective measure of performance, we categorized them into 58 categories, mainly according to their respective artists. One artist corresponds to one category. We did not distinguish those artists for antique Arab, Egyptian, Chinese, and India paintings and just put them into four categories.

- *Online survey on 642 painting images.* We collected 642 painting images from Internet, mainly impressionism paintings and modern arts from 30 artists. To evaluate the algorithm performance on completely true user preferences, we performed a web-based online survey[5] to gather user ratings for 642 images. In the survey, each user gave ratings, i.e. "like", "dislike", or "not sure", to a randomly selected set of painting images. We so collected data from more than 200 visitors. After removing users who had rated less then 5 images, and users who had rated all of their images with one class (only like resp. only dislike), we retain a total of $L = 190$ users. On average, each of them had rated 89 images.

In both settings, we extract and combine *color histogram* (216-dim.), *correlagram* (256-dim.), *first and second color moments* (9-dim.) and *Pyramid wavelet texture* (10-dim.) to form 491-dimensional feature vectors to represent images. Across all the experiments, we use SVMs with RBF (radius basis function) kernel. In our empirical study, we will mainly examine the accuracy of collaborative ensemble learning in terms of predicting users' interests in art images, and compare it with other two competitive algorithms:

- *SVM content-based retrieval* trains a SVM model on a set of examples given by an active user, and then apply the model to predict the active user's preferences. This algorithm represents a typical CBIR approach.

- *Collaborative filtering* combines a society of advisory users' preferences to predict an active user's preferences. The combination is weighted by *Pearson correlation* between test user and other advisory users' preferences. The algorithm applied in this paper is described in [4].

4.1 Simulation with 4533 Painting Images

In this study, we will examine the retrieval accuracy of collaborative ensemble learning in cases that users have heterogenous interests for art images based on the 4533 painting images.

To enable objective evaluation, we need to "mimic" many users' preferences for the images. We assume that each user is interested in n categories. Since painting images from the same artist (e.g. one category) typically share similar painting styles, the assumption reflects the real-world cases to some extent, where one is interested in heterogeneous styles of paintings. We further assume that, without loss of generality, for a setting of n, there is $\mathcal{P}_n$, a set of profile types containing $58-n+1$ profile types and the p-th profile type is interested in n adjacent categories from the p-th to the $(p+n-1)$-th one.[6] Then we stimulate a user's preference data in the following steps:

1. Randomly choose the value of n, where n can be 1, 2, or 3. Each possibility has equal chance.
2. Randomly assign a profile type in $\mathcal{P}_n$ to the user, where each profile type has equal chance.
3. Randomly produce 5 liked art images and 10 disliked art images based on the profile type assigned.

We repeat the procedure 1000 times and thus produce 1000 users' preference data. The detailed setting-up is based on some assumptions, however, we believe that it approaches real-world cases from certain perspectives. Since it is not easy to gather the *ground truth*, i.e. sufficient true-user preferences for an art image base, it is necessary to perform simulations at this early stage.

Our experiments take a leave-one-out scheme, in which a user is picked up as a test user (i.e. active user) and the remaining ones serve as *advisory users*. Then the test user's profile type serves as the ground truth for evaluation. Based on the profile type, we generate a number of examples, with approximately 1/3 liked images and 2/3 disliked ones, to feed the art image retrieval system. We use top-N accuracy to measure the performance, i.e. the fraction of truly liked images among the N top ranked images. We change the number of given examples for each active user, i.e. 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50, to study the learning curve of the three compared methods. For one learning curve, we repeat the procedure for 10 times with different random seeds and each run will go through all the active users. Finally we compute the mean and standard deviation of the mean over

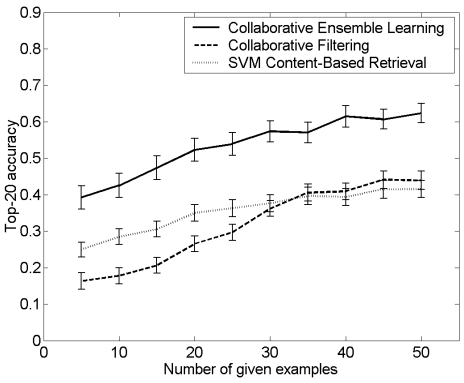

Figure 5: Top-20 accuracy with various number of given examples for each active user. For each advisory user, we assume that 5 liked and 10 disliked images are given (Simulation on 4533-painting data)

[5]The survey can be found on `http://honolulu.dbs.informatik.uni-muenchen.de:8080/paintings/index.jsp`.

[6]The image categories are sorted in alphabet order of artist names.

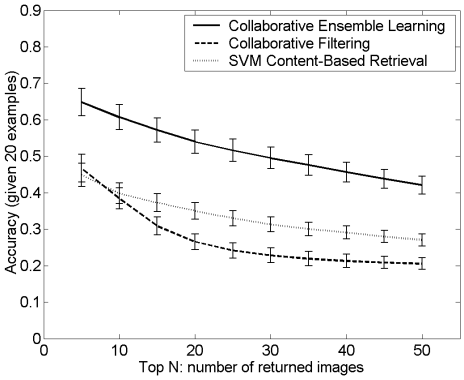

Figure 6: Accuracy with various number of returned images. For each active user, we fix the number of given examples to 20. For each advisory user, we assume that 5 liked and 10 disliked images are given (Simulation on 4533-painting data)

the 10 runs. The obtained final results have been shown in Fig. 5. Collaborative ensemble learning significantly outperforms the other two algorithms, which indicates that the algorithm effectively captures simulated users' diverse interests for art images. While the SVM content-based retrieval shows a poor accuracy. The results confirm our analysis that although SVM demonstrate excellent learning performances in many real-world problems, it suffers the problems of modelling users' diverse interests due to the deficiency of low-level features. Collaborative filtering performs the worst in our simulation, because the preference ratings given by advisory users are very sparse, i.e. only 0.33% of the images are rated for each user. Collaborative filtering heavily relies on the user ratings while ignoring the descriptive features of images. It cannot compute reliable Pearson correlation between two users if they have few commonly rated examples. While our proposed collaborative ensemble learning generally overcomes the weaknesses of SVM content-based approach and collaborative filtering by incorporating wider information and thus achieves the best accuracy.

In the following, we fix the number of given examples for each active user to 20 and vary N, the number of top ranked results that are returned. Accuracy is then computed for all the active users and the procedure is repeated for 10 times with different random seeds. Fianlly the mean and error bar of the mean are calculated and demonstrated in Fig. 6. Accuracies of the three approachs are all decreasing as we increase the number of N, indicating that all the three methods present ranking which is better than random guess (which should be a flat line with accuracy insensitive to the value of N). However, collaborative ensemble learning clearly demonstrates the best performance. Interestingly, collaborative filtering's accuracy decreases the most quickly with N increasing. This is because that collaborative filtering is not able to generalize examples to similar cases (i.e. images distributed very close to the given examples in the low-level feature space), and thus cannot make judgements on the images never visited by any adisory user (i.e. new images). Therefore, it "consumes out" those limited number of liked images which could be suggested by advisory users at the early stage and cannot present more

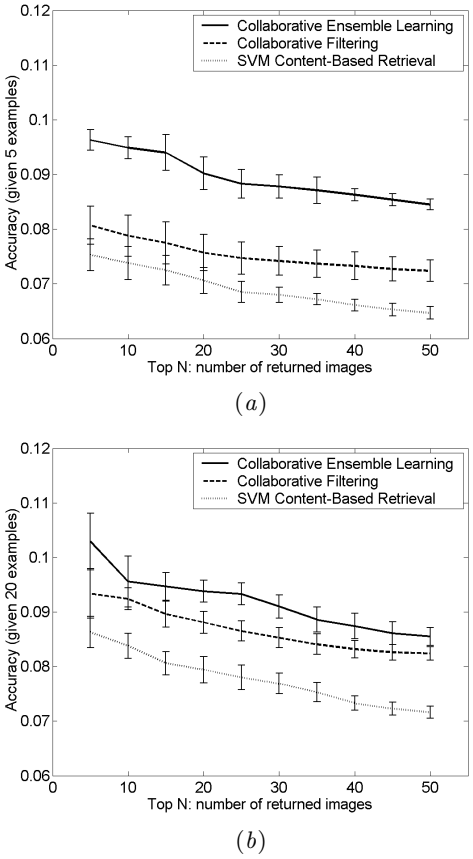

Figure 7: Accuracy with various number of returned images. (a) for each active user, we assume that 5 examples are given, (b) for each active user, we assume that 20 examples are given

truly liked images when N futher increases. This observation indicates that content-based approach has the ability of generalizing examples to never-rated cases, and clearly collaborative ensemble learning takes over and further enhances this advantage.

4.2 Experiments with the Online Survey Data

Although we get impression that collaborative ensemble learning presents excellent performances, however, simulation can not replace the real-world cases. In this section, we will examine the performance of the three approaches based on 190 user's preference data on 642 painting images, which are gathered from the on-line survey. Again, we use top-N accuracy to evaluate the performance. Since we can not require a user to rate all of the 642 painting images in the surey, for each user we just partially know the "ground truth" of preferences. As a result, the true precision cannot be computed. We thus adopt the accuracy measure that is the fraction of *known* liked images in top ranked N images. The quantity is smaller than true accuracy because *unknown* liked images are missing in the measurement. However, in our survey, the presenting of images to users is completely random, thus the distributions of rated/unrated images in both unranked and ranked lists are also random. This randomness dose not change the relative values of compared

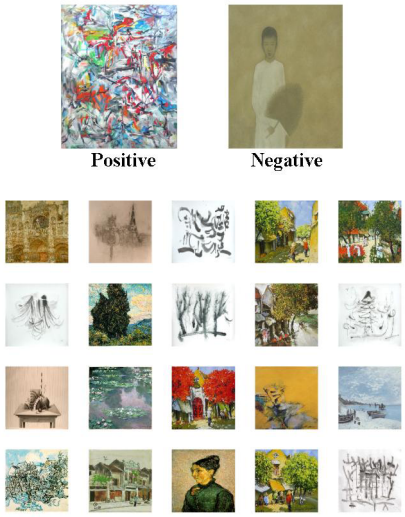

Figure 8: Case study: Two images on the top are examples given by a user. The lower 20 images are the top-20 results returned by collaborative ensemble learning.

methods but just the absolute values. Thus in our following experiment it still makes sense to use the adopted accuracy measurement to compare the three retrieval methods.

Our experiment takes the leave-one-out scheme again, in which we pick up each user as the active user and treat all other users as collected advisory users. We fix the number of given examples for each active user to 5 and 20 respectively, and examine the retrieval accuracy in the cases of returning various N top ranked images. We take the same methodology as Fig. 6 and demonstrate the results in Fig. 7-(a) (given 5 examples) and Fig. 7-(b) (given 20 examples). We find that collaborative ensemble learning achieves the best accuracy in both cases. Since in the data user ratings are much denser than the simulation case, collaborative filtering outperforms the SVM content-based method. Interestingly, the accuracy improvement of collaborative ensemble learning over the other two approaches are more impressive in the given-5 case. This is a very nice property for art image retrieval because users are normally not patient at the initial information-gathering stage and it is much desired to get satisfactory accuracy with only a few examples. Theoretically, this nice property can be explained from the Bayesian perspective (Sec. 3.1), where we use "an informative prior" learned from all the users to constraint the Bayesian inference. Such a prior knowledge gained from population promises a good accuracy even when limited examples are fed to the learning system.

In the next, we take a closer look at a case study. As shown in Fig. 8, we let a user input a positive and a negative examples to run the collaborative ensemble learning algorithm. The returned top 20 results look quite diverse and meanwhile very different from the positive example. Surprisingly, the user loves 18 out of the 20 images and there is no strongly disliked image. As a comparison, we present the results of SVM content-based approach trained on the same examples in Fig. 9. We find that 8 results are actually from the same artist as the positive example is. The user

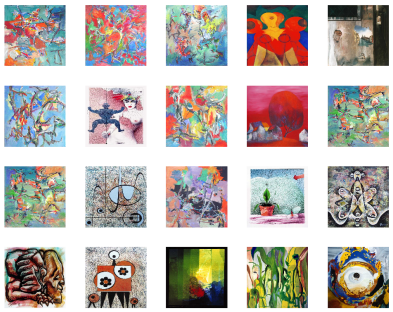

Figure 9: Top-20 results returned by SVM content-based retrieval. Examples are the same as the ones shown in Fig. 8.

told us that he strongly dislikes the images (1,4), (3,2) (3,5), (4,1), (4,3), (4,4) and (4,5).[7] This case study is quite interesting, which demonstrates that, in the studied case where a user gives examples that only partially convey his preferences, collaborative ensemble learning effectively infer the user's comprehensive interests while SVM content-based approach only returns images that are similar to the positive example(s). In the art image retrieval application, presenting interesting but *novel* images to active users is a very nice property because a user can easily find images from the same artist (by category-based search) while has difficulties in locating potentially interesting images which are currently unknown to the user.

5. RELATED WORK

5.1 Image Retrieval

There have been extensive studies on content-based image retrieval (CBIR). All the investigated approaches approximately fall into two categories, i.e. similarity-based and learning-based approaches. A straightforward way of CBIR is to measure the similarity of images to the given examples. Research on this family of methods has been centered around the following issues: (1) extraction of image features like color, texture, region, and hybrid features, e.g. [27, 18, 6], (2) query reweighting and query center movement [15, 19, 23], and (4) efficient similarity search [15].

Image retrieval can be cast as a machine learning problem. Given some relevant and irrelevant example images from a user, a classifier is trained and applied to classify all the unseen images. One important advantage of learning-based methods is the effective means to model complex distributions. Moreover, learning-based approaches opened opportunities for further improvements of CBIR systems, like relevance feedback[34], active learning [29, 7] and transductive learning [31]. However, a common limitation of existing algorithms is that they only consider the limited examples given by current query user while ignoring the examples given by others. Some efforts were made to overcome this problem, which infer a semantic space from historical user relevance feedbacks (e.g. [11]). Recently, ensemble learning gained many attentions in machine learning community,

[7]Here we treat the presented 20 images as a 4 by 5 matrix.

like boosting [10] and bagging [5]. Tieu and Viola made one early attempt to apply boosting to image retrieval[28].

5.2 Collaborative Filtering

A variety of CF algorithms have been proposed in the last decade. The earliest memory-based algorithms were based on the observation that people usually trust the recommendations from like-minded friends, like the movie recommender systems described in [24, 26]. Many newly proposed CF methods fall into the class of model-based CF and are inspired from machine learning algorithms. Examples include linear classifiers [33], Bayesian networks [4], dependency networks [12] and latent-class models[14]. Pure CF only mines user experiences (e.g. scores, clicks, purchases) while dose not incorporates content of items. It greatly suffers from the extreme sparsity of data and the new-item problem. To overcome these weaknesses, some efforts were recently made to combine CF with content-based filtering, however, mainly in an *ad-hoc* way where the results of two methods are weighted averaged [1, 20, 2]. There are only few examples of a unifying framework for these two basic information filtering ideas, one being the three-way aspect model of [22], which is only applicable to text data. Our work unifies CF and content-based filtering in a sound Bayesian framework and, in principle, can be applied to any media, including image, text and video.

6. CONCLUSIONS AND FUTURE WORK

To our best knowledge, this paper made one of the earliest attempts to study the problem of art image retrieval (AIR)[8]. Based on analysis of properties of AIR and existing possible solutions, i.e. CF and CBIR, we present a statistical approach, named collaborative ensemble learning, to meet the challenges of AIR task. The algorithm builds profiling models for each user based on low-level visual features (the CBIR idea), and uses weighted combination of many models to make predictions for active users (the CF idea). Our experiments based on two data sets demonstrate that the described algorithm is significantly superior over SVM-based CBIR algorithm and Pearson-correlation CF algorithm, in terms of prediction accuracy.

The major advantages of our work are (1) A principled framework unifying CBIR and CF has been introduced to solve the problems in art image retrieval; (2) Since theoretically any probabilistic models (not only PSVMs) can also be adopted, collaborative ensemble learning can be flexibly tailored to various tasks, like video or music retrieval; (3) Being built on the top of normal content-based retrieval models, the described approach makes one step futher by allowing "communications" among many related but not identical models when making predictions. This nice property suggests that it can be easily integrated into existing content-based media retrieval systems and adapts them to be sensitive to user preferences.

However, there are still some limitations in this paper, which will motivate our future work. (1) We need to further reduce the time and memory costs of collaborative ensemble learning. A possible solution can be selecting users with typical profiles thus the model is working on a small number of L; (2) The PSVM model are restricted with two-class examples, i.e. users must provide both liked and disliked images. As indicated in Sec. 3.2, the collaborative ensemble learning is generally applicable to any probabilistic models other than SVMs. It might be interesting to try other density models within this framework; (3)As pointed out by two anonymous reviewers, the current work only focuses on "prediction" rather than "explanation", namely, it lacks the power to further study which style of paintings (e.g. impressionism and abstract impressionism) might be of interest to which group of people (e.g. general public, curator, and historian). One direction of future work is to extend our current model with explanatory abilities, for example, grouping users based on their preferences or grouping images based on associated users, and conduct a more comprehensive user survey for getting data that support the corresponding evaluation.

[8]One other research project on art image retrieval was performed by James Wang at Pennsylvania State University. (For details please visit http://art.ist.psu.edu/.)

7. ACKNOWLEDGMENTS

This work was partially done when the first author was visiting Microsoft Research Asia at Beijing. We would thank the three anonymous reviewers and Dr. Nevenka Dimitrova for their constructive comments to this work. We would also thank Franz Krojer for his efforts in maintaining the on-line survey at University of Munich.

8. REFERENCES

[1] M. Balabanovic and Y. Shoham. Fab: Content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72, 1997.

[2] C. Basu, H. Hirsh, and W. W. Cohen. Recommendation as classification: Using social and content-based information in recommendation. In *Proceedings of the Fifteenth National Conference on Artificial Intelligencen AAAI/IAAI*, pages 714–720, 1998.

[3] C. M. Bishop, M. Svensen, and C. K. Williams. GTM: The generative topographic mapping. *Neural Compuation*, 10(1):215–234, 1998.

[4] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 43–52, 1998.

[5] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

[6] E. Chang, B. Li, and C. Li. Toward perception-based image retrieval. In *IEEE Content-Based Access of Image and Video Libraries*, pages 101–105, June 2000.

[7] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Pichunter: Bayesian relevance feedback for image retrieval. In *Proceedings of International Conference on Pattern Recognition*, volume 3, pages 361–369, Austria, 1996.

[8] G. Fishman. *Monte Carlo Concepts, Algorithms and Applications*. Springer Verlag, 1996.

[9] M. Flickher, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D.Steele, and P. Yanker. Query by image and video content: The qbic system. *IEEE Computer*, 28(9):23–32, 1995.

[10] Y. Freund and R. Schapire. A short introduction to boosting. *J. Japan. Soc. for Artif. Intel.*, 14(5):771–780, 1999.

[11] X. He, W.-Y. Ma, O. King, M. Li, and H. Zhang. Learning and infering a semantic space from user's relevance feedback for image retrieval. In *Proceedings of ACM conference on Multimedia*, 2002.

[12] D. Heckerman, D. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.

[13] G. Hinton, C. Williams, and M. Revow. Adaptive elastic models for hand-printed character recognition. In J. Moody, S. Hanson, and R. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, pages 512–519. Morgan Kauffmann, 1992.

[14] T. Hofmann and J. Puzicha. Latent class models for collaborative filtering. In *Proceedings of IJCAI'99*, pages 688–693, 1999.

[15] Y. Ishikawa, R. Subramanya, and C. Faloutsos. Mindreader: Querying databases through multiple examples. In *VLDB*, 1998.

[16] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *Proceedings of European Conference on Machine Learning*. Springer, 1998.

[17] W.-Y. Ma and B. Manjunath. Netra: A toolbox for navigating large image database. *ACM Mulitimedia Systems*, 7:184–198, 1999.

[18] B. S. Manjunath and W.-Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[19] M. Ortega, Y. Rui, K. Chakrabarti, A. Warshavsky, S. Mehrotra, and T. Huang. Supporting ranked boolean similarity queries in mars. *IEEE Trans. on Knowledge and Data Engineering*, 10(6):905–925, December 1999.

[20] M. Pazzani. A framework for collaborative, content-based and demographic filtering. *Artificial Intelligence Review*, 13(5–6):393–408, 1999.

[21] J. C. Platt. Probabilities for SV machines. In A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74, Cambridge, MA, 1999. MIT Press.

[22] A. Popescul, L. Ungar, D. Pennock, and S. Lawrence. Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *17th Conference on Uncertainty in Artificial Intelligence*, pages 437–444, Seattle, Washington, August 2–5 2001.

[23] K. Porkaew, K. Chakrabarti, and S. Mehrotra. Query refinement for multimedia similarity retrieval in mars. In *Preceedings of ACM Multimedia*, November 1999.

[24] P. Resnick, N. Iacovou, M. Sushak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 Computer Supported Collaborative Work Conference*, pages 175–186. ACM, 1994.

[25] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Trans. Circuits and Systems for Video Tech.*, 8(5):644–655, 1998.

[26] U. Shardanand and P. Maes. Social information filtering algorithms for automating 'word of mouth'. In *Proceedings of ACM CHI'95 Conference on Human Factors in Computing Systems*, volume 1, pages 210–217, 1995.

[27] J. Smith and S.-F. Chang. Automated image retrieval using color and texture. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, November 1996.

[28] K. Tieu and P. Viola. Boosting image retieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 228–235, 2000.

[29] S. Tong and E. Chang. Support vector machine active learning for image retrieval. In *Proceedings of ACM conference on Multimedia*, pages 107–118, Ottawa, Canada, 2001.

[30] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, New York, 1995.

[31] Y. Wu, Q. Tian, and T. Huang. Discriminant-em algorithm with application to image retrieval. In *Proc. of IEEE Conf. on CVPR'2000*, volume I, pages 222–227, 2000.

[32] K. Yu, A. Schwaighofer, V. Tresp, W.-Y. Ma, and H. Zhang. Collaborative ensemble learning: Combining collaborative and content-based information filtering via hierarchical bayes. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2003.

[33] T. Zhang and V. S. Iyengar. Recommender systems using linear classifiers. *Journal of Machine Learning Research*, 2:313–334, 2002.

[34] X. Zhou and T. Huang. Exploring the nature and variants of relevance feedback. In *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries, in conjunction with CVPR01*, Hawaii, 2001.

[35] X. Zhou and T. Huang. Small sample learning during multimedia retrieval using biasmap. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, Hawaii, 2001.

APPENDIX

A. TUNING THE PARAMETERS OF PSVMS

The RBF kernel parameters as well as the constant C, are chosen to minimize the leave-one-out error on the training data. Since the training set for most users is very small, this typically leads to over-fitting. Thus, the kernel parameters and C are shared among models, and the optimization is with respect to the average leave-one-out error on all models.

For choosing the slope A_i of the sigoidal function Eq. (3), we follow the three-fold crossvalidation strategy suggested by [21]. The training data are divided into three equally sized subsets. An SVM model is trained on two of these subsets and evaluated on the third subset as a test set. The SVM outputs on all three test sets are stored, we denote them by $f^{3\mathrm{CV}}(\boldsymbol{x}_j)$. As the final step, the slope A_i is chosen such as to maximize the regularized log-likelihood of data in the union of the three test sets, using the SVM outputs found in the cross validation procedure.