Hindawi Publishing Corporation
EURASIP Journal on Advances in Signal Processing
Volume 2007, Article ID 83526, 10 pages
doi:10.1155/2007/83526

Research Article

A Learning State-Space Model for Image Retrieval

Cheng-Chieh Chiang,[1,2] Yi-Ping Hung,[3] and Greg C. Lee[4]

[1] *Department of Information and Computer Education, College of Education, National Taiwan Normal University, Taipei 106, Taiwan*
[2] *Department of Information Technology, Takming College, Taipei 114, Taiwan*
[3] *Graduate Institute of Networking and Multimedia, College of Electrical Engineering and Computer Science, National Taiwan University, Taipei 106, Taiwan*
[4] *Department of Computer Science and Information Engineering, College of Science, National Taiwan Normal University, Taipei 106, Taiwan*

Received 30 August 2006; Accepted 12 March 2007

Recommended by Ebroul Izquierdo

This paper proposes an approach based on a state-space model for learning the user concepts in image retrieval. We first design a scheme of region-based image representation based on concept units, which are integrated with different types of feature spaces and with different region scales of image segmentation. The design of the concept units aims at describing similar characteristics at a certain perspective among relevant images. We present the details of our proposed approach based on a state-space model for interactive image retrieval, including likelihood and transition models, and we also describe some experiments that show the efficacy of our proposed model. This work demonstrates the feasibility of using a state-space model to estimate the user intuition in image retrieval.

1. INTRODUCTION

Image retrieval has become a very active research area since the 1990s due to the rapid increase in the use of digital images [1, 2]. Estimating the user concepts is one of the most difficult tasks in image retrieval. Feature extraction involves extracting only low-level features such as color, texture, and shape from an image. However, people understand an image semantically, rather than via the low-level visual features, and there is a large gap between the low-level features and the high-level concepts in image understanding [3].

The relevance feedback approach [4, 5] is widely used for bridging this semantic gap. In each iteration of a retrieval task, the user assigns some relevant and irrelevant examples according to their concepts, from which the system learns to estimate what the user actually wants. Many types of learning models have been applied in relevance feedback for image retrieval, such as Bayesian framework [6–8], SVM [9], and active learning [10]. Goh et al. also proposed several quantitative measures to model concept complexity in the learning of relevance feedback [10].

Image representation is another important issue that needs to be addressed when solving the above problem. It is necessary to design good units for image representation even if a perfect learning approach is applied to image retrieval. Many recent studies have adopted the region-based approach [9, 11, 12] for image representation, because region features can be more representative for user requests than global image features. Constructing a set of visual words [13, 14] that collects similar region features to be a representative unit is appropriate for region-based image representation. Image annotation [15, 16] is another method that labels an image with high-level information. Some researchers have attempted to build a semantic space for describing the high-level concepts in images [17, 18].

In this paper, we present a new scheme for image representation and propose a learning model for image retrieval. Instead of constructing a fixed semantic space for representing the user concepts, we have designed a flexible scheme based on *concept units* for region-based image representation that combines different types of feature spaces and different scales of image segmentation. We also propose an interactive approach for estimating the user concepts implicit in the user feedbacks in a query session, which is the period between when the first query is made to when the corresponding relevance feedbacks are produced. Our basic idea is to

track the behaviors of the user concepts of relevance feedbacks in image retrieval using a state-space model [19–21]. The state-space model has been well defined and widely applied to dynamic systems. However, we did not find studies in the literature that have applied the state-space model to the learning problem in relevance feedback. Our work aims at demonstrating the feasibility of solving the retrieval problem using a state-space model.

This paper is organized as follows. Section 2 introduces the motivation and the idea behind our proposed approach. Section 3 describes the proposed concept units used in region-based image representation, and the proposed learning model based on a state-space model is shown in Section 4. Section 5 presents the image ranking method used to determine the similarity of two images. Section 6 describes a strategy for handling negative examples. Section 7 details some experiments that applied our approach, and Section 8 draws conclusions and discusses future work.

2. MOTIVATION

We consider the problem of category search in image retrieval. This involves grouping images into the same category that the user perceives to be semantically relevant. For example, the image set from Corel Photo, a set of image data widely used in many researches, contains many types of semantic categories. Hence we consider a user called "Corel Photo" who chooses relevant images to form these categories. Note that different users may assign different semantic categories in the same image set. The main challenge for category search is to estimate the user concepts, for example, Corel Photo, from the interaction of the retrieval.

Let a query session comprise the first query and corresponding relevance feedbacks. We assume that the user does not change the requesting concepts, that is, the semantic concepts in a query session are constant. Ideally, we can view the process of obtaining relevance feedbacks as tracing the path from the first query to the retrieval goals, from which we can estimate the user concepts in a retrieval task.

During a retrieval task, the user could have a semantic goal but could be unable to describe it explicitly—the retrieval target exists but is not explicit in the beginning of the retrieval. For example, the user may want to retrieve images of flowers but will be unable to describe their types wanted until she/he looks at relevant images. For this scenario, we can model the tracing path of the user concepts as

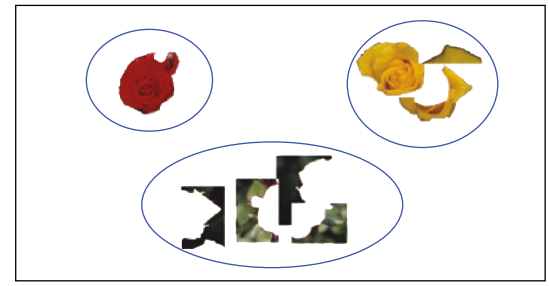
$$X_1 = IM \cdot X_{t-1} + \eta_{t-1}, \tag{1}$$

where X_t means the user state at the tth iteration, IM is the identical matrix, and η_{t-1} is the noise term (i.e., variations of user concepts in relevance feedbacks). We estimate each stage of the tracing path using the state X_t, which is determined from the previous estimated states and various types of feedbacks specified by the user.

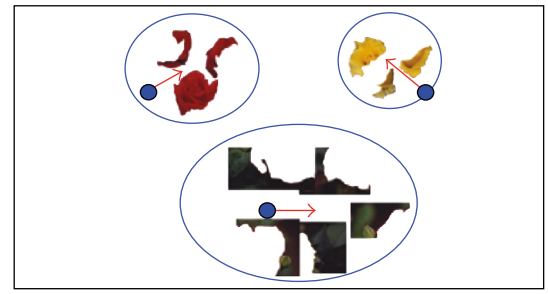
Figure 1 illustrates our idea that tracks the relevant region features in the feature space to estimate the user concepts in image retrieval. Figures 1(a) and 1(b) show the two sets of relevant images that are specified by the user at tth and $(t+1)$th iterations, respectively. Figures 1(c) and 1(d) describe the process of tracking the movement of relevant regions in a visual feature space. At tth iteration, it is assumed that the relevant region features involve three components shown in Figure 1(c). Hence we can depict these region features using the centroids (i.e., means) of the three components. At the next iteration, the estimation of the state starts with the previous centroids, drawn as blue dots in Figure 1(d), and moves to the current relevant regions.

(a) Relevant examples at the tth iteration

(b) Relevant examples at the $(t+1)$th iteration



(c) Region features at the tth iteration



(d) Tracking the movement of region features from tth iteration to $(t+1)$th iteration

Figure 1: An illustration of tracking the movement of region features in relevance feedbacks.

In this work, we aim at solving (1) to estimate the user concepts relevant to image retrieval. We assume that state X_t can be modeled using a Gaussian mixture [22] with means $\boldsymbol{\mu}_t$ and variances $\boldsymbol{\sigma}_t$, where $\boldsymbol{\mu}_t$ represent the user concepts in state X_{t-1}, and $\boldsymbol{\sigma}_t$ are the variances of the user feedbacks in noise term η_{t-1}. In the example of Figure 1, a pair of $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ forms a blue dotted circle to represent the user concept at

an iteration. Solving means $\boldsymbol{\mu}_t$ and variances $\boldsymbol{\sigma}_t$ requires two major tasks: representation and estimation for the state.

We first have to design a scheme for representing the state, which intuitively handles the semantic gap between visual features and user concepts. We do not try to directly construct a semantic space for image retrieval because it is impossible to explicitly describe what the user wants before requests are made. In this work, we design a flexible scheme using concept units that are based on combinations of different types of region features and different scales of image segmentation. Any two images that are designated as relevant by the user should be similar from a certain perspective. The concept units are designated to represent unknown perspectives of relevant images based on the user perceptions.

We next design an iterative approach for learning and estimating the user state. The idea of estimating the tracing path of relevance feedbacks motivated us to design a state-space model of the user state described in (1). The state-space model has been widely applied to analyze and infer dynamic systems according to information on time sequences. In our proposed model, the time sequence for the state-space model is associated with the iteration process of relevance feedbacks, and the training data for learning or inferring the system is extracted from positive examples in the relevance feedbacks. Moreover, we design a simple strategy for handling negative examples in order to eliminate false alarms in retrieval results.

3. CONCEPT UNITS FOR REGION AND IMAGE REPRESENTATION

3.1. Image segmentation and feature extraction

Region-based approach is widely used to the analysis of image contents. To extract regions, the first task is to partition an image into multiple regions using image segmentation. The most intuitive method for image segmentation is to segment objects (or foreground subjects) for region-based image matching [9, 11–13]. However, this is very difficult, and the segmentation results greatly affect the performance of region-based tasks. Hence, some researchers have divided an image into rectangular girds [15] or a large number of overlapping circular regions [23].

Generally speaking, image segmentation may not be consistent with human perception. Our proposal is not to generate the perfect regions with segmentation, but rather to determine useful ones. We use the well-known watershed segmentation [24], which is an efficient, automatic, and unsupervised segmentation method for gray-level images, to partition an image into nonoverlapping regions. A color image is first converted to a gray image and then partitioned by the watershed segmentation. A watershed region is often homogeneous in the intensity space, and that means that pixels in a watershed region are not very diverse. Hence, the watershed regions are appropriate for representing the region units of an image. Wang proposed a multiscale approach for watershed segmentation in order to overcome the problem of oversegmentation [24], which is the major drawback of the original method of watershed segmentation, by controlling the scaling parameters. Different scaling parameters result in different numbers of regions being segmented in the same image.

Assume that the database contains N images, denoted as $\{I_1, \ldots, I_N\}$, and that v scales, denoted as $S = \{s_1, \ldots, s_v\}$, are used for watershed segmentation. Given a scale s_q, we assume there are n_q regions to be partitioned for all images in the database. Thus, we can annotate the set of regions as

$$\{r_1^{s_q}, \ldots, r_{n_q}^{s_q}\}. \tag{2}$$

Let the set of features $F = \{f_1, \ldots, f_u\}$ contain u different types of visual features. Given a feature type f_p, the feature vector extracted from region $r_i^{s_q}$ is written as $f_p(r_i^{s_q})$. Thus, given a feature type f_p and a scale s_q, we have a set of feature vectors, denote that as R_p^q, with respect to the set of watershed regions in (2):

$$R_p^q = \bigcup_{i=1}^{n_q} f_p(r_i^{s_q}), \quad 1 \le p \le u,\ 1 \le q \le v. \tag{3}$$

Note that the region representation described above is independent of selecting visual features and segmentation methods. We collect different scales and different features of regions for an image in order to represent unknown perspectives of relevant images. Using more types of visual features and more scales of regions covers a wider range of the image contents, but makes the computational complexity excessive. In this work, four types of visual features (i.e., $u = 4$) are used: (i) color histogram, (ii) color moments (both color features are in HSV space), (iii) cooccurrence texture, and (iv) Gabor texture. Moreover, we set $v = 2$, that is, two types of region scales, in the watershed segmentation.

3.2. Concept units

Since it is impossible to predict the best way to represent an image, for example, which type of features or which scale for image segmentation is better for image representation, before the user makes the query, we first collect different types of region representation, and then estimate which is best for characterizing the user's perceptions in relevance feedbacks. R_p^q, in (3), represents the collection of visual features of watershed regions that are observed using different scales and different features, hence giving a total of $u \times v$ types of region features with$v$ scaling parameters and u types of visual features.

Given the feature type f_p and the scaling parameter s_q, we apply the K-means algorithm [22] to cluster the feature vectors R_p^q. That is, we partition the feature space into K areas. Suppose $C_p^q(1), \ldots, C_p^q(K)$ are the clusters for all regions with respect to s_q and f_q. Collecting all of the region features yields the clusters:

$$\bigcup_{p,q} \bigcup_{k=1}^{k} C_p^q(k). \tag{4}$$

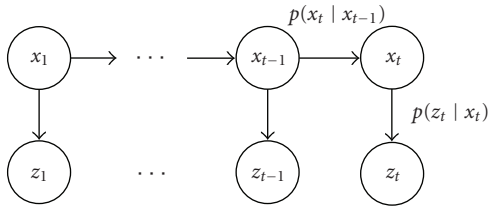


Figure 2: The probabilistic structure of the state-space model.

These $u \times v \times K$ clusters are the concept units for all $1 \le p \le u$, $1 \le q \le v$, and $1 \le k \le K$ representing images in the entire image database with different scalings and different features. The definition of concept units is a variant of the so-called visual word [13, 14], which draws the processing units in the space of the visual features. The generation of the concept units with different types of feature spaces and with different region scales provides more possibilities to fit the different characteristics of the image contents for semantically relevant images. In our experiments, we set K at 400, hence giving $u \times v \times K = 4 \times 2 \times 400 = 3\,200$ concept units.

3.3. Region-based image representation

We can build the concept units in (4) for all images in the database in order to represent the types of contents that the user retrieves. Therefore, we design a region-based image representation based on the concept units. Let I be an image in the database. For each concept unit $C_p^q(k)$, where $1 \le p \le u$, $1 \le q \le v$, and $1 \le k \le K$, let the weight $w_p^q(k)$ be the ratio of the number of regions belonging to $C_p^q(k)$ to the number of regions in image I. Thus, we collect all weights, $w_p^q(k)$, to from a $(u \times v \times K)$-dimensional vector for representing image

$$\{w_p^q(k) \mid 1 \le p \le u,\ 1 \le q \le v,\ 1 \le k \le K\}. \tag{5}$$

4. LEARNING MODEL BASED ON A STATE-SPACE MODEL

4.1. State-space model

The state-space approach has been widely applied to the analysis of dynamic systems, which involve estimating the state of a system which changes over time from a sequence of noisy measurements [19]. Many papers have detailed state-space models [19–21], and hence here we only provide a brief summary of how the posterior probability of a state-space model is inferred.

Figure 2 depicts the probabilistic structure of the Bayesian network of a state-space model, which contains two types of nodes at time t: (i) x_t for the system state and (ii) z_t for the observation measurement. At time t, the dynamic system receives inputs z_t, for which we want to estimate the posterior probability of the system state x_t given the past observations; this is denoted as $p(x_t \mid z_{1,\dots,t})$, where $z_{1\dots t}$ represents the collection of observations z_1 to z_t. Two assumptions are generally applied to a state-space model for simplicity. The first is the first-order Markov property, given by

$$p(x_t \mid x_{1,\dots,t-1}) = p(x_t \mid x_{t-1}), \tag{6}$$

where $x_{1,\dots,t-1}$ represents the collection of states x_1 to x_{t-1}. The second is that the observations are mutually independent:

$$p(z_t \mid x_t, z_{1\dots,t-1}) = p(z_t \mid x_t), \tag{7}$$

where $z_{1,\dots,t-1}$ means the collection of the observations z_1 to $z_t - 1$. By using the above two assumptions and Bayes' rule, the posterior probability of state x_t given the past observations can be inferred as

$$p(x_t \mid z_{1,\dots,t}) = \frac{p(z_t \mid x_t)\, p(x_t \mid z_{1,\dots,t-1})}{p(z_t \mid z_{1,\dots,t-1})}, \tag{8}$$

where

$$p(x_t \mid z_{1,\dots,t-1}) = \sum_{x_{t-1}} p(x_t \mid x_{t-1})\, p(x_{t-1} \mid z_{1,\dots,t-1}). \tag{9}$$

Thus, we can infer the posterior probability as

$$\begin{aligned} &p(x_t \mid z_{1,\dots,t}) \\ &\quad = \frac{p(z_t \mid x_t)}{p(z_t \mid z_{1,\dots,t-1})} \sum_{x_t - 1} p(x_t \mid x_{t-1})\, p(x_{t-1} \mid z_{1,\dots,t-1}) \\ &\quad \propto p(z_t \mid x_t) \sum_{x_{t-1}} p(x_t \mid x_{t-1})\, p(x_{t-1} \mid z_{1,\dots,t-1}). \end{aligned} \tag{10}$$

In (10), the posterior probability $p(x_t \mid z_{1,\dots,t})$ in a state-space model is recursively based on two factors: (i) a system model $p(x_t \mid x_{t-1})$ which describes the evolution of the state over time (called the *transition function*), and (ii) a measurement model $p(z_t \mid x_t)$ which relates the observation and noise to the state (called the *observation function*). It is also necessary to define the prior probability of state $p(x_1)$ at the beginning of the recursion.

4.2. The proposed learning model

The user intuition is usually implicit in the specification of positive and negative examples in the query session. Positive examples are generally used to estimate the user intuition, and negative examples are used as exceptions in the estimation. Hence, we apply the positive examples of the tth iteration of relevance feedbacks to observations z_t of the tth stage of the state-space model, and the negative examples are used to eliminate the false alarms in retrieval results. The strategy for handling the negative examples is described in Section 6.

The user concepts X_t, stated in (1), can be approximated by a Gaussian mixture model with means $\boldsymbol{\mu}_t$ and variances $\boldsymbol{\sigma}_t$ where the means $\boldsymbol{\mu}_t$ indicate the concept units for representing the user concepts, and the variances $\boldsymbol{\sigma}_t$ cover the varying scopes of the user concepts in the concept units. Intuitively, the state vector for the state-space model could be defined as a set of the pairs of means and variances for the Gaussian

mixture model. However, this makes the model very complex, and also we do not have a huge training data set for learning and inferring the model because the number of positive examples is not large in a query session. Hence, it is necessary to simplify the design of the state-space model for image retrieval.

In this work, we simplify the definition of the state vector in two ways. The first is to ignore the variances $\boldsymbol{\sigma}_t$. The definition of concept units covers some variances because they are defined as clusters in the feature space. Ignoring the variances $\boldsymbol{\sigma}_t$ in defining the state vector means that we assume that the variance of concepts is limited to the radius of the concept units. The second is to define a single concept unit which is viewed as a greedy method instead of multiple concept units in the state vector. Considering the tth iteration in a query session, let x_t be the most representative concept unit for the user concepts that we want to estimate, and let $z_{1,\ldots,t}$ be the collection of positive examples of relevance feedbacks. Thus, we want to find the maximal posterior estimation of state x_t given the past positive examples (observations $z_{1,\ldots,t}$) in relevance feedbacks:

$$x_t^* = \arg\max p(x_t \mid z_{1,\ldots,t}). \tag{11}$$

The user concepts in the query session generally comprise multiple rather than single factors, and hence we take the first H highest probabilities of x_t^* to represent the user concepts.

Below we define the state vector, observation function, and transition function that are used to construct the state-space model.

State vector

We define the state as the most representative concept unit for the query session. The definition of concept unit $C_p^q(k)$ is associated with feature type p, region scale q, and cluster k, and thus we define the state vector as a three-dimensional vector denoted as (p, q, k), where $1 \le p \le u$, $1 \le p \le v$, and $1 \le k \le K$.

Observation function

Let the positive images of relevance feedbacks be the observations of the state-space model. We define the observation function $p(z_t \mid x_t)$ as the likelihood of the observation given each state,

$$p(z_t \mid x_t) = \frac{\text{no. of computed concept units in positive images}}{\text{no. of all concept units in positive images}}. \tag{12}$$

Let us consider an example in which there are 100 regions in relevant images at an iteration of a query session. Therefore, these observations contain 100 concept units because each region feature belongs to a concept unit. If 35 regions fall in the same concept unit, its observation measurement is $35/100 = 0.35$.

Transition function

The transition model $p(x_t \mid x_{t-1})$ is designed to model the variations of concept units representing the user concepts in iterations of relevance feedbacks. The transition function must record the changing cost between any two concept units. Given two state vectors $v_1 = (p_1, q_1, k_1)$ and $v_2 = (p_2, q_2, k_2)$ with $p_1 \neq p_2$, this means that the two units are from different feature spaces. Because different types of features capture different characteristics in images, it is inappropriate to estimate the state cross-different features. Hence we set the transition function $\text{Trans}(v_1, v_2)$ to 0 if $p_1 \neq p_2$. We next consider the case in which concept units are in the same feature space, that is, $p_1 = p_2$. Thus, we can compute the meaningful distance between these two concept units either with or without the same region scale. However, the transition measurement of concept units crossing different scales should be less than that in the same scale. Let $M(p_1, q_1, p_2, q_2)$ be a $K \times K$ matrix in which each element M_{ij} is the Euclidean distance between concept units (p_1, q_1, i) and (p_2, q_2, j). Note that M_{ij} corresponds to the Euclidean distance between the means of clusters $C_{p_1}^{q_1}(i)$ and $C_{p_2}^{q_2}(j)$. We then define the transition function as

$$\text{Trans}\left(v_1(p_1, q_1, k_1), v_2(p_2, q_2, k_2)\right) = \begin{cases} \dfrac{2 \cdot \exp(-M_{k_1 k_2})}{\sum_y \exp(-M_{k_1 y})} & \text{if } p_1 = p_2,\ q_1 = q_2, \\ \alpha \cdot \dfrac{2 \cdot \exp(-M_{k_1 k_2})}{\sum_y \exp(-M_{k_1 y})} & \text{if } p_1 = p_2,\ q_1 \neq q_2, \\ 0 & \text{if } p_1 \neq p_2, \end{cases} \tag{13}$$

where α is a scaling factor with $0 \le \alpha \le 1$. Note that $\alpha = 0.5$ in our implementation.

Prior distribution

All of the prior probabilities of the states are set equal. This means that the tracking of the model starts at all concept units.

At the beginning of the iterations, all concept units have equal probabilities for representing the query concepts. During the process of relevance feedbacks in the query session, representative concept units from observations will have higher probabilities based on the inference of the state-space model using (10). We take first H concept units with maximal posterior probabilities to represent the user concepts at each iteration.

Two factors are involved in image retrieval based on the proposed state-space model: (i) the likelihoods of positive examples and (ii) the transitive conditions between any two concept units. The former is commonly applied in a Bayesian framework, and the latter is not common in image retrieval. An interesting approach to the transition is to use the ontological structure which represents a domain of knowledge in image retrieval [25, 26]. Note that embedding these two factors in relevance feedbacks is one of the main contributions of our proposed model.

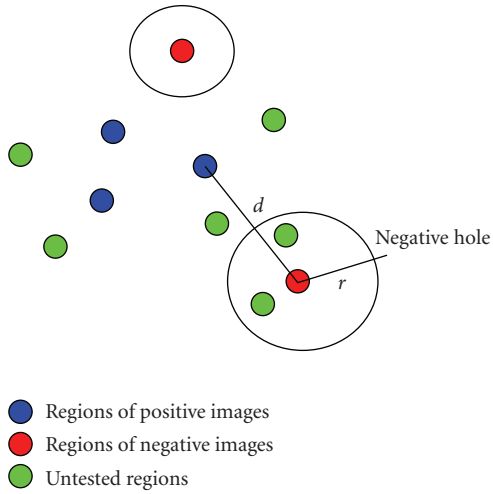


Figure 3: An illustration of the negative holes, d: distance to the nearest positive region, r: the radius of the negative hole, $d/2$.

5. IMAGE RANKING

The proposed learning model uses H concept units to largely represent the concepts the user retrieves in a query session. A similarity measure between the retrieval concepts and an image in the database is used for image matching and ranking. Without loss of generality, let the first H concept units with maximal posterior probabilities at the tth iteration be denoted by $v_{\tau(i)}$, where $1 \le i \le H$. The posterior probabilities of these H concepts are described by

$$p_t(i) = p\left(x_t\left(v_{\tau(i)}\right) \mid z_t\right), \quad 1 \le i \le H, \tag{14}$$

where $\tau(i)$ is the index of concept units, and $x_t(v_{\tau(i)})$ is the state with concept unit $v_{\tau(i)}$ at the tth iteration.

The idea of designing the similarity measure is to find images containing most of the H concept units in (14). Since an image I in the database can be represented as (5), we design a dissimilarity measure between the retrieval concepts of the query session and the image I at the tth iteration as follows:

$$\mathrm{DisSim}(I,t) = \left(\sum_{i=1}^{M}\left(w_{\tau(i)} - p_t(i)\right)^2\right)^{1/2} \tag{15}$$

6. STRATEGY FOR HANDLING NEGATIVE EXAMPLES

The previous sections only use positive examples of feedbacks for learning the concepts that the user wants to retrieve. While negative examples could be applied in the learning model to decrease the rate of false retrieval results, handling them is difficult because they are diverse either in feature spaces or in semantic concepts. In our opinion, a negative example only removes some of the false retrieval results in a localized area. In this work, we adopt the strategy following from [27] for handling negative examples. The basic idea is to excavate a "negative hole" in the feature space around the regions of each negative example. Figure 3 illustrates an example of negative holes. The center of a negative hole is a region feature of a negative image, and its radius is half the distance from the negative region to the nearest positive one. Each iteration of relevance feedbacks involves the generation of many negative holes associated with regions of negative examples. A region of a test image in the database is neglected in computing weights $w_p^q(k)$ in (5) if it falls in a negative hole.

7. EXPERIMENTAL RESULTS AND DISCUSSION

7.1. Dataset

In our experiments, we used three datasets (denoted as DI, DII, and DIII) where DI and DII contain photo images collected from Corel Photo and DIII is Caltech-101 Object Categories [28].

Dataset DI

DI contains 20 categories and each category consists of 100 photo images. All images can be partitioned into over 70 000 regions with two scales of image segmentation. These images contain a wide range of contents, such as landscapes, animals, plants, and buildings. These data categories are classified according to human concepts such as "beautiful rose," "autumn," and "doors in Paris," and hence even images in the same category may have had diverse contents. However, all images in the same category are viewed as relevant to each other.

Dataset DII

We extended DI to the larger dataset DII which contains 50 categories, each consisting of 100 photo images, giving a total of 5 000 images. All images can be partitioned into over 200 000 regions with two scales of image segmentation. For each category in DI and DII, we randomly choose 10 images as the query, so the size of the query set is 200 and 500 images, respectively. Moreover, 10 iterations are performed for each query.

Dataset DIII

We took the Caltech-101 Object Categories [28] as the third dataset that is publicly available and involves 101 categories of objects with over 8 000 images. The number of images in each category is different. Over 300 000 regions are segmented with two scales of image segmentation. We randomly chose 10 images as the query for the larger categories which contain more than 80 images, giving a total of 240 query images.

7.2. Evaluation and discussion

The precision and the recall are commonly used to evaluate the performance of a retrieval system. Note that precision $= A/B$ and recall $= A/C$, where A is the number of relevant images that we retrieve, B is the number of returned images in the retrieval, and C is the number of all relevant images

Table 1: The detailed precisions using DI without handling negative examples.

Cat ID	$t = 1$	$t = 2$	$t = 3$	$t = 4$	$t = 5$	$t = 6$	$t = 7$	$t = 8$	$t = 9$	$t = 10$
0	0.354	0.497	0.549	0.556	0.557	0.558	0.559	0.559	0.559	0.559
1	0.134	0.251	0.305	0.332	0.349	0.352	0.355	0.355	0.355	0.355
2	0.154	0.302	0.398	0.432	0.443	0.447	0.453	0.457	0.457	0.457
3	0.156	0.273	0.381	0.446	0.479	0.491	0.493	0.495	0.496	0.496
4	0.177	0.268	0.378	0.485	0.531	0.548	0.553	0.554	0.554	0.554
5	0.241	0.475	0.633	0.713	0.752	0.754	0.758	0.758	0.758	0.758
6	0.247	0.404	0.548	0.651	0.705	0.722	0.724	0.725	0.726	0.726
7	0.156	0.266	0.386	0.484	0.538	0.555	0.555	0.555	0.556	0.565
8	0.245	0.428	0.547	0.583	0.606	0.607	0.608	0.609	0.613	0.634
9	0.415	0.644	0.782	**0.849**	**0.883**	**0.884**	**0.884**	**0.884**	**0.884**	**0.884**
10	0.221	0.395	0.497	0.533	0.543	0.545	0.546	0.562	0.641	0.709
11	0.285	0.548	0.657	0.672	0.673	0.673	0.673	0.693	**0.810**	**0.859**
12	0.205	0.352	0.455	0.504	0.521	0.524	0.539	0.556	0.730	0.788
13	0.223	0.375	0.464	0.513	0.523	0.531	0.563	0.701	0.760	0.798
14	0.238	0.358	0.496	0.593	0.643	0.667	0.724	**0.823**	**0.895**	**0.919**
15	0.297	0.484	0.576	0.592	0.633	0.743	**0.876**	**0.893**	**0.893**	**0.893**
16	0.450	0.611	0.752	**0.847**	**0.888**	**0.912**	**0.959**	**0.967**	**0.968**	**0.968**
17	0.216	0.386	0.537	0.612	0.712	**0.833**	**0.888**	**0.888**	**0.888**	**0.888**
18	0.283	0.461	0.602	0.668	0.736	**0.851**	**0.883**	**0.887**	**0.890**	**0.890**
19	0.197	0.312	0.444	0.568	0.695	**0.838**	**0.874**	**0.888**	**0.888**	**0.889**
AVG	0.245	0.404	0.519	0.582	0.620	0.652	0.673	0.690	0.716	0.730

($C = 100$ in DI and DII). We set $B = 100$ in our system, hence precision = recall in datasets DI and DII. Moreover, some of the categories contain more than 100 images in dataset DIII. Thus, we employ the recall instead of the precision to evaluate the performance of the proposed method in our experiments.

Figure 4 shows the average recalls at each iteration of relevance feedbacks in five cases: only using DI without handling negative examples, and using DII and DIII with/without handling negative examples. DI-pos exhibits the highest recalls because the size of DI is smaller than that of DII and DIII. However, the performances of DII-pos+neg and DIII-pos+neg indicate that handling negative example can significantly improve the retrieval.

Table 1 lists the detailed recalls of all categories of DI of relevance feedbacks using our proposed model without negative examples. The first row in Table 1 denotes the iteration of relevance, and the last row indicates the average precisions of all image categories. Note that precisions larger than 0.8 are shown in boldface.

Both Figure 4 and Table 1 indicate that the retrieval performances are bad at the beginning of the retrieval. The reason is that only few positive feedbacks at the beginning are available, and hence the training data are insufficient for accurately estimating the states. After several iterations, the efficacy of the proposed model is more manifest.

We now discuss the experiments in detail. Figures 5 and 6(b) illustrate two cases that correspond to better and worse retrieval results, respectively, using DII without handling negative examples. Figure 5(a) shows some images of the categories "bus" and "butterfly" for which our proposed model produces better results, and Figure 5(b) lists the average precisions of the two categories at each iteration. Similarly, Figure 6(a) shows example images of the categories

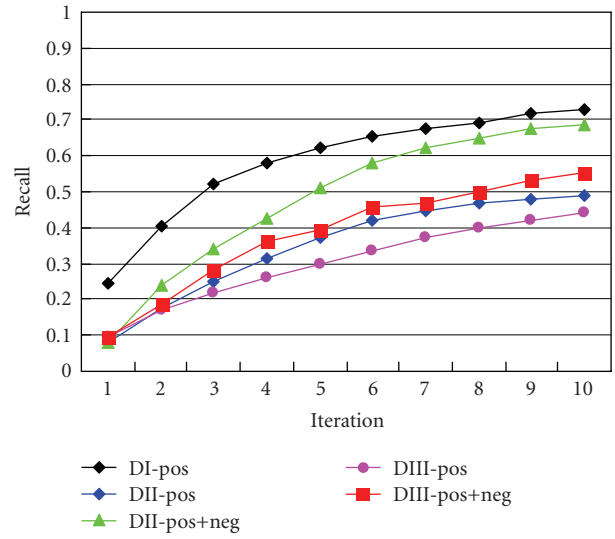


Figure 4: Average recalls for the three datasets DI-pos, DII-pos, and DIII-pos: using these datasets without handling negative examples; DII-pos+neg, and DIII-pos+neg: using the two datasets with handling negative examples.

(a) The first and second rows are examples of categories "bus" and "butterfly," respectively

Cat.	1	2	3	4	5	6	7	8	9	10
Bus	0.179	0.316	0.437	0.543	0.658	0.758	0.824	0.863	0.878	0.896
Butt.	0.067	0.122	0.175	0.222	0.39	0.704	0.782	0.81	0.938	0.969

(b) The detailed precisions of the categories "bus" and "butterfly," respectively

Figure 5: Illustrations of better results using DII without handling negative examples.

(a) The first and second rows are examples of categories "in desert" and "snow mountain," respectively

Cat.	1	2	3	4	5	6	7	8	9	10
Des.	0.057	0.09	0.118	0.151	0.178	0.19	0.193	0.194	0.194	0.194
Snow	0.048	0.09	0.116	0.146	0.151	0.17	0.18	0.186	0.188	0.188

(b) The detailed precisions of the categories "in desert" and "snow mountain," respectively

Figure 6: Illustrations for worse results using DII without handling negative examples.

"in desert" and "snow mountain" that have worse results, and Figure 6(b) shows their average precisions. In the better cases of Figure 5, images in the same category have the same semantic concepts but still look quite different. This shows the feasibility of using the proposed approach to model images with similar semantic concepts but diverse visual features. However, huge variations either in visual features or semantic concepts are still very difficult to model. For example, the "snow mountain" images in Figure 6 are easily confused with those in other landscape categories.

Basically, our approach is appropriate for image retrieval with relevance feedbacks. The time sequences in the state-space model can be easily associated with the iterations of relevance feedbacks. The proposed model does not only involve the likelihoods of positive images, but also considers the transition possibilities among concept units. However, two problems are worth solving in our approach. The first is the smaller number of positive examples at the beginning of the feedbacks. This is a common problem in image retrieval because no users enjoy manually assigning a huge number of positive examples in the feedback process. One method for solving this problem is to design a long-term strategy to include all positive examples of previous query sessions as training data. The second problem is the huge variations between images in the same category. A possible method for solving this problem is to make our model more complex by embedding more information. However, this could result in overfitting, especially since we do not have many training data in relevance feedbacks. Constructing a knowledge structure such as the ontology-based approach [25, 26] is

potential in image retrieval if the retrieval task focuses on an application domain. After defining the transition model of the structure for the knowledge domain, our proposed model can consider both the low-level features (likelihood model) and high-level concepts (transition model) for bridging the semantic gap problem in image retrieval.

8. CONCLUSIONS AND FUTURE WORK

This work demonstrates the feasibility of solving the problem of the semantic gap for image retrieval using a state-space model. We design concept units, which integrate with different types of visual features and with different scales of image segmentation, for image representation. We also propose a state-space model for estimating the user concepts in a query session. Our approach involves both the likelihood model of positive examples and the transition model among concept units in image retrieval. Moreover, we have presented a strategy for handling negative feedbacks for refining the retrieval results in this paper.

Some future tasks are required to extend this work. The first is to define a long-term learning strategy for solving the problem of a small training set at the beginning iterations of relevance feedbacks. The second is to integrate the knowledge structure for a domain application with the transition model in our proposed approach. Moreover, the design of concept units could be revised to contain higher-level information rather than visual features. Other methods of machine learning, such as active leaning or boosting, could be integrated with the state-space model for image retrieval.

ACKNOWLEDGMENTS

This work was supported in part by the Ministry of Economic Affairs, Taiwan, under Grant 95-EC-17-A-02-S1-032 and by the Excellent Research Projects of National Taiwan University under Grant 95R0062-AE00-02.

REFERENCES

[1] R. Datta, J. Li, and J. Z. Wang, "Content-based image retrieval: approaches and trends of the new age," in *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval (MIR '05)*, pp. 253–262, Singapore, November 2005.

[2] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: state of the art and challenges," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 2, no. 1, pp. 1–19, 2006.

[3] K. Goh, B. Li, and E. Y. Chang, "Semantics and feature discovery via confidence-based ensemble," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 1, no. 2, pp. 168–189, 2005.

[4] Y. Rui, T. S. Huang, M. Ortega, and S. Mehrotra, "Relevance feedback: a power tool for interactive content-based image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 5, pp. 644–655, 1998.

[5] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: a comprehensive review," *Multimedia Systems*, vol. 8, no. 6, pp. 536–544, 2003.

[6] I. J. Cox, M. L. Miller, T. P. Minka, T. V. Papathomas, and P. N. Yianilos, "The Bayesian image retrieval system, PicHunter: theory, implementation, and psychophysical experiments," *IEEE Transactions on Image Processing*, vol. 9, no. 1, pp. 20–37, 2000.

[7] Z. Su, H. Zhang, S. Li, and S. Ma, "Relevance feedback in content-based image retrieval: Bayesian framework, feature subspaces, and progressive learning," *IEEE Transactions on Image Processing*, vol. 12, no. 8, pp. 924–937, 2003.

[8] N. Vasconcelos and A. Lippman, "Learning from user feedback in image retrieval systems," in *Proceedings of Advances in Neural Information Processing Systems (NIPS '99)*, pp. 977–986, Denver, Colo, USA, November-December 1999.

[9] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "An efficient and effective region-based image retrieval framework," *IEEE Transactions on Image Processing*, vol. 13, no. 5, pp. 699–709, 2004.

[10] K.-S. Goh, E. Y. Chang, and W.-C. Lai, "Multimodal concept-dependent active learning for image retrieval," in *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pp. 564–571, New York, NY, USA, October 2004.

[11] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: a system for region-based image indexing and retrieval," in *Proceedings of the 3rd International Conference on Visual Information and Information Systems (VISUAL '99)*, pp. 509–516, Amsterdam, The Netherlands, June 1999.

[12] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: semantics-sensitive integrated matching for picture libraries," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, pp. 947–963, 2001.

[13] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *Proceedings of the 8th IEEE International Conference on Computer Vision (ICCV '01)*, vol. 2, pp. 408–415, Vancouver, BC, Canada, July 2001.

[14] L. Fei-Fei and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*, vol. 2, pp. 524–531, San Diego, Calif, USA, June 2005.

[15] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple Bernoulli relevance models for image and video annotation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '04)*, vol. 2, pp. 1002–1009, Washington, DC, USA, June-July 2004.

[16] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03)*, pp. 119–126, Toronto, Ont, Canada, July-August 2003.

[17] D. R. Heisterkamp, "Building a latent semantic index of an image database from patterns of relevance feedback," in *Proceedings of the 16th International Conference on Pattern Recognition (ICPR '02)*, vol. 4, pp. 134–137, Quebec, Canada, August 2002.

[18] A. Shah-Hosseini and G. M. Knapp, "Learning image semantics from users relevance feedback," in *Proceedings of the 12th Annual ACM International Conference on Multimedia*, pp. 452–455, New York, NY, USA, October 2004.

[19] M. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *IEEE Transactions on Signal Processing*, vol. 50, no. 2, pp. 174–188, 2002.

[20] Z. Ghahramani, "An introduction to hidden Markov models and Bayesian networks," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 15, no. 1, pp. 9–42, 2001.

[21] K. P. Murphy, *Dynamic Bayesian networks: representation, inference and learning*, Ph.D. thesis, University of California, Berkeley, Calif, USA, 2002.

[22] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, New York, NY, USA, 2nd edition, 2001.

[23] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman, "Learning object categories from Google's image search," in *Proceedings of the 10th IEEE International Conference on Computer Vision (ICCV '05)*, vol. 2, pp. 1816–1823, Beijing, China, October 2005.

[24] D. Wang, "A multiscale gradient algorithm for image segmentation using watersheds," *Pattern Recognition*, vol. 30, no. 12, pp. 2043–2052, 1997.

[25] V. Mezaris, I. Kompatsiaris, and M. G. Strintzis, "An ontology approach to object-based image retrieval," in *Proceedings of IEEE International Conference on Image Processing (ICIP '03)*, vol. 2, pp. 511–514, Barcelona, Spain, September 2003.

[26] M. Srikanth, J. Varner, M. Bowden, and D. I. Moldovan, "Exploiting ontologies for automatic image annotation," in *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '05)*, pp. 552–558, Salvador, Brazil, August 2005.

[27] I. Atmosukarto, W. K. Leow, and Z. Huang, "Feature combination and relevance feedback for 3D model retrieval," in *Proceedings of the 11th International Multimedia Modelling Conference (MMM '05)*, pp. 334–339, Melbourne, Australia, January 2005.

[28] L. Fei-Fei, R. Fergus, and P. Perona, "Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories," in *Proceedings of IEEE CVPR Workshop of Generative Model Based Vision (WGMBV '04)*, Washington, DC, USA, June 2004.

Cheng-Chieh Chiang received his B.S. degree in applied mathematics from Tatung University, Taipei, Taiwan, in 1991, and his M.S. degree in computer science from National Chiao Tung University, Hsinchu, Taiwan, in 1993. He is currently working towards the Ph.D. degree in Department of Information and Computer Education, National Taiwan Normal University, Taipei, Taiwan. His research interests include multimedia information indexing and retrieval, pattern recognition, machine learning, and computer vision.

Yi-Ping Hung received his B.S. degree in electrical engineering from the National Taiwan University in 1982. He received his M.S. degree from the Division of Engineering, his M.S. degree from the Division of Applied Mathematics, and his Ph.D. degree from the Division of Engineering, all at Brown University, in 1987, 1988, and 1990, respectively. He is currently a Professor in the Graduate Institute of Networking and Multimedia, and in the Department of Computer Science and Information Engineering, both at the National Taiwan University. From 1990 to 2002, he was with the Institute of Information Science, Academia Sinica, Taiwan, where he became a Tenured Research Fellow in 1997 and is now an Adjunct Research Fellow. He served as a Deputy Director of the Institute of Information Science from 1996 to 1997, and received the Young Researcher Publication Award from Academia Sinica in 1997. He has served as the Program Cochair of ACCV'00 and ICAT'00, as the Workshop Cochair of ICCV'03, and as a member in the editorial board of the International Journal of Computer Vision since 2004. His current research interests include computer vision, pattern recognition, image processing, virtual reality, multimedia, and human-computer interaction.

Greg C. Lee received his B.S. degree from Louisiana State University in 1985, and his M.S. and Ph.D. degrees from Michigan State University in 1988 and 1992, respectively, all in computer science. Since 1992, he has been with the National Taiwan Normal University where he is currently a Professor at the Department of Computer Science and Information Engineering. His research interests are in the areas of image processing, video processing, computer vision, and computer science education. Dr. Lee is a Member of IEEE and ACM.