MULTI-LAYER SEMANTIC REPRESENTATION LEARNING FOR IMAGE RETRIEVAL

Wei Jiang, Guihua Er, Qionghai Dai

Broad Band Network & Digital Media Lab, Dept. Automation, Tsinghua University, Beijing, China

ABSTRACT

Long-term relevance feedback learning is an important learning mechanism in content-based image retrieval. In this paper, our work has two contributions: (1) A *Multi-layer Semantic Representation* (*MSR*) is proposed, and an algorithm is implemented to automatically build the MSR for image database through long-term relevance feedback learning. (2) The accumulated MSR is incorporated with the short-term feedback learning to help subsequent users' retrieval. The MSR memorizes the multi-correlation among images, and integrates these memories to build hidden semantic concepts for images, which are distributed in multiple semantic layers. In experiment, an MSR is built based on the real retrieval from 10 different users, which can precisely describe the hidden concepts underlying images, and help to bridge the gap between high-level concepts and low-level features, and thus improve the retrieval performance significantly.

1. INTRODUCTION

Content-based image retrieval (*CBIR*) has been largely explored since last decades. In a CBIR system, images the user wants usually share some semantic cue, which is called the *hidden semantic concept* underlying the images. The gap between high-level hidden concepts and low-level visual features has been the main obstacle for the development of CBIR systems, and the *relevance feedback* technique is introduced to bridge this gap [8]. There are two kinds of relevance feedback mechanisms. The *short-term learning* can be viewed as a supervised learning process [7]. During a query session, the user labels some images as "relevant" or "irrelevant" to the query concept in each feedback round, and supervises the system to adjust search in subsequent retrieval rounds. However the learned information is discarded when a new query session begins. The *long-term learning* [1, 3, 4, 5, 6] memorizes the information gained from user's feedback during previous query sessions, and utilizes the accumulated experience to help retrieval in subsequent query sessions, which usually has better performance than only using the short-term learning.

There are mainly three long-term learning mechanisms. First, the *Latent Semantic Indexing* (*LSI*) approach [4] memorizes the "relevant" images for each query to form image indexes for query concepts. Second, the bi-correlation approach [6] records statistical bi-correlation between each image pair to calculate the semantic similarity between images, which is combined with low-level similarity to help retrieval. Third, the clustering approach clusters the images into several clusters, with each cluster representing one hidden semantic concept: in [1, 3] images are initially clustered by low-level features and the feedback information is used to adjust the clustering results; in [5] images are clustered by semantic similarities from bi-correlation records. From the aspect of concept learning, the LSI and bi-correlation approaches explicitly don't learn semantic concepts from the long-term records, while the clustering approach contains further learning mechanism to extract semantic clusters and reveal hidden semantic concepts.

These approaches can exploit the accumulated experience in some sense to help retrieval. However, the real world hidden semantics have the following two characteristics: (1) Instead of bi-correlation between images, the query concept usually can only be reflected by multi-correlation among a group of images. For example images in Fig.1 come from the semantic category "fruit". When (a, d, e) are labeled as "relevant", and others "irrelevant", the query concept is "orange"; when (a, c, f) are "relevant", and others "irrelevant", the query concept is "green color fruit". (2) Instead of the one kind of hard partition in one semantic layer divided by clustering method, the real world concepts should have multiple *semantic layers*, with one layer corresponding to one kind of hard partition of hidden semantic space. Some concepts have intrinsic *intersections*, e.g., "green color" and "orange" describe objects' different attributes. Image sets belonging to these two concepts can't be hard divided, and should be in different semantic layers.

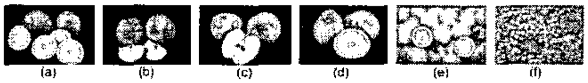


Fig. 1. Example for semantic exposition by group of images.

To address the issue of long-term feedback learning with real semantics extraction, our work in this paper has two contributions. (1) A *Multi-layer Semantic Representation* (*MSR*) for image database is proposed and an algorithm is implemented to automatically build the MSR through long-

Thanks to CNNSF (60172026) for funding.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

Authorized licensed use limited to: Columbia University. Downloaded on December 29, 2008 at 22:41 from IEEE Xplore. Restrictions apply.

term relevance feedback learning process. The MSR records the direct multi-correlation among images, and extracts hidden concepts, which are distributed in multiple semantic layers, from these memories. One semantic layer corresponds to one kind of hard partition of semantic space. Concepts without intersection between each other are put into one semantic layer, and concepts with intersection are put into different layers. (2) An integrated algorithm is proposed to seamlessly combine the semantic information provided by accumulated MSR with the short-term learner to help retrieval in subsequent query sessions. The experiment is carried by totally 1000 rounds of real retrieval from 10 different users, which shows that the MSR built can describe the real world semantic concepts underlying images precisely. Also, the incorporation of long-term MSR and short-term learning can help to grasp the query concept, and improve the retrieval performance significantly.

2. MULTI-LAYER SEMANTIC REPRESENTATION

2.1. MSR formulation

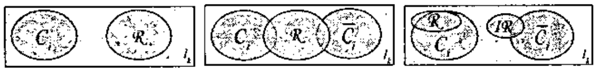
Assume that we have already learned N semantic concepts $c_1, \ldots, c_N$, which are distributed in M semantic layers $l_1, \ldots, l_M$. $\mathcal{C}_i$ and $\overline{\mathcal{C}}_i$ denote the corresponding "relevant" and "irrelevant" image sets for c_i, respectively, where image $\mathbf{x} \in \mathcal{C}_i$ belongs to concept c_i, and image $\mathbf{x} \in \overline{\mathcal{C}}_i$ is labeled to be "irrelevant" with c_i. If $c_i \in l_k$ and $c_j \in l_k$, $\mathcal{C}_i \cap \mathcal{C}_j = \phi$. Also, we want the number of layers to be as small as possible, and thus each layer to be as full as possible.

Suppose in a new query session, images in $\mathcal{R}$ are labeled to be "relevant", and images in $\mathcal{IR}$ "irrelevant" by user, c_q denotes the current query concept. Let Ct_i be the relationship between c_q and existing c_i, and Lt_k be the relationship between c_q and existing l_k. We have following definitions.

Definition 1 [Ct_i]:

$$Ct_i = \begin{cases} -1, & \mathcal{R} \cap \mathcal{C}_i = \phi \\ 0, & \mathcal{R} \cap \mathcal{C}_i \neq \phi, \text{ and } \mathcal{R} \cap \overline{\mathcal{C}}_i \neq \phi \text{ or } \mathcal{IR} \cap \overline{\mathcal{C}}_i \neq \phi \\ 1, & \mathcal{R} \cap \mathcal{C}_i \neq \phi \text{ and } \mathcal{R} \cap \overline{\mathcal{C}}_i = \phi \text{ and } \mathcal{IR} \cap \overline{\mathcal{C}}_i = \phi \end{cases}$$

Where $Ct_i = -1$ means that c_q and c_i have no relationship (Fig.2(a)); $Ct_i = 0$ means that c_q and c_i are *incompatible* (can't be in the same layer) (Fig.2(b)); $Ct_i = 1$ means c_q is *compatible* with c_i (Fig.2(c)).



(a) $Ct_i = -1$ (b) $Ct_i = 0$ (c) $Ct_i = 1$

Fig. 2. Examples for relationship between c_q and existing c_i.

Definition 2 [Lt_k]: Set $n_0^k = \sum_{c_i \in l_k} I(Ct_i = 0)$, $n_1^k = \sum_{c_i \in l_k} I(Ct_i = 1)$, where $I(A) = 1$ if A is true, $I(A) = 0$ else.

$$Lt_k = \begin{cases} 0, & n_0^k > 0 \text{ or } n_1^k > 1 \\ 1, & \text{otherwise} \end{cases}$$

$Lt_k = 0$ means $c_q \notin l_k$ (Fig.3(a)); $c_q \in l_k$ if $Lt_k = 1$ (Fig.3(b)).

$\mathcal{C}_i$
$\mathcal{R}$
$\mathcal{C}_j$
l_k
$\mathcal{C}_i$
$\mathcal{R}$
$\mathcal{C}_j$
l_k

(a) $Lt_k = 0$ (b) $Lt_k = 1$

Fig. 3. Examples for relationship between c_q and existing l_k.

Let S_L and S_C denote the layer id and concept id of the current query concept, c_q, respectively. S_L and S_C determine the relation between c_q and the existing semantics, and can be learned by algorithm in Fig.4. Thus the MSR for the database can be adaptively built by the algorithm in Fig.5.

```
Algorithm: Semantic Status Learning
Input: $l_1, \ldots, l_M, \mathcal{C}_1, \ldots, \mathcal{C}_N, \overline{\mathcal{C}}_1, \ldots, \overline{\mathcal{C}}_N, \mathcal{R}, \mathcal{IR}$
Output: Layer id $S_L$, Concept id $S_C$
    1 $n_t = \sum_{k=1}^{M} I(Lt_k = 1)$; // number of layers which
                                  // are compatible with $c_q$
    2 If ($n_t = 0$) { $S_L = M+1$, $S_C = N+1$; } // new concept
                                                  // in new layer
      Else if ($n_t = 1$) { $S_L = \arg_k\{Lt_k = 1\}$;
          If ($n_1^k = 1$) { $S_C = \arg_i\{Ct_i = 1, c_i \in l_k\}$; }
          Else { $S_C = N+1$;} // new concept in $l_k$
      }
      Else { $m_t = \sum_{Lt_k=1} n_1^k$; // the number of compatible
                                // concept in compatible layers
          If ($m_t = 0$) {// new concept in the lowest layer
              $S_L = \arg\min_k\{Lt_k = 1\}$; $S_C = N+1$; }
          Else if ($m_t = 1$) { $S_L = \arg_k\{n_1^k = 1\}$;
                          $S_C = \arg_i\{Ct_i = 1,\ c_i \in l_k\}$; }
          Else { // the semantic status of $c_q$ is not sure
              $S_L = 0$; $S_C = 0$; }
      }
```

Fig. 4. Pseudo-code for semantic status learning.

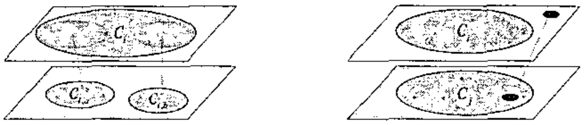
```
Algorithm Long-Term MSR Learning:
Input: $S_L$, $S_C$
If ($S_L = M+1$) {create $l_{M+1}$, $c_{N+1} = c_q$, $c_{N+1} \in l_{M+1}$;}
Else if ($1 \leq S_L \leq M$) {
    If ($1 \leq S_C \leq N$) { $\mathcal{C}_{S_C} = \mathcal{C}_{S_C} \cup \mathcal{R}$; $\overline{\mathcal{C}}_{S_C} = \overline{\mathcal{C}}_{S_C} \cup \mathcal{IR}$;}
    Else { create $c_{N+1} = c_q$, $c_{N+1} \in l_k$; } }
Else { don't record this query; }
```

Fig. 5. Pseudo-code for long-term MSR learning.

2.2. Post processing

There are two situations where a concept may be wrongly built: (1) Images belonging to concept $\mathcal{C}_i$ have a great variety in low-level features, and are learned into different $\mathcal{C}_{i,a}$ and $\mathcal{C}_{i,b}$. When more query sessions are carried, the true $\mathcal{C}_i$ is built, but in another layer (Fig.6 (a)); (2) Some images belonging to $\mathcal{C}_j$ is mislabeled as in $\overline{\mathcal{C}}_i$, and form $\mathcal{C}_i$ in another layer (Fig.6 (b)) (the mislabeling problem is common for all CBIR systems). The post processing is necessary to alleviate these mistakes and increase the robustness of our method. Here we adopt the concept merging method.

Authorized licensed use limited to: Columbia University. Downloaded on December 29, 2008 at 22:41 from IEEE Xplore. Restrictions apply.



(a) (b)

Fig. 6. Two kinds of wrongly extracted concepts.

For an n size database, viewing images $\mathbf{x}_1, \dots, \mathbf{x}_n$ as attributes for semantic concepts, vector $\mathbf{F}(\mathcal{C}_i) = [I(\mathbf{x}_1 \in \mathcal{C}_i), \dots, I(\mathbf{x}_n \in \mathcal{C}_i)]$ and $\mathbf{F}(\overline{\mathcal{C}}_i) = [I(\mathbf{x}_1 \in \overline{\mathcal{C}}_i), \dots, I(\mathbf{x}_n \in \overline{\mathcal{C}}_i)]$ can be viewed as the binary feature vectors for set $\mathcal{C}_i$ and $\overline{\mathcal{C}}_i$. The similarity between concepts c_i and c_j can be measured by the famous *Feature Contrast Model* (*FCM*) [2] (which is a successful psychological similarity measurement) as:

$$S(\mathcal{C}_i, \mathcal{C}_j) = Sim(\mathcal{C}_i, \mathcal{C}_j) - Dis(\mathcal{C}_i, \mathcal{C}_j) - Dis(\mathcal{C}_j, \mathcal{C}_i) \quad (1)$$

where $Sim(\mathcal{C}_i, \mathcal{C}_j)$ and $Dis(\mathcal{C}_i, \mathcal{C}_j)$ are given by:

$$Sim(\mathcal{C}_i, \mathcal{C}_j) = \frac{\mathbf{F}(\mathcal{C}_i) \cdot \mathbf{F}(\mathcal{C}_j)}{\min\{||\mathbf{F}(\mathcal{C}_i)||^2,\ ||\mathbf{F}(\mathcal{C}_j)||^2\}}$$

$$Dis(\mathcal{C}_i, \mathcal{C}_j) = \frac{\mathbf{F}(\mathcal{C}_i) \cdot \mathbf{F}(\overline{\mathcal{C}}_j)}{\min\{||\mathbf{F}(\mathcal{C}_i)||^2,\ ||\mathbf{F}(\overline{\mathcal{C}}_j)||^2\}}$$

Where $\cdot$ is the dot product. When $S(\mathcal{C}_i, \mathcal{C}_j) > \alpha$, $\mathcal{C}_i$ and $\mathcal{C}_j$ are merged. α is statistically set to be 0.9 in our system.

3. RETRIEVAL BY BOTH LONG-TERM AND SHORT-TERM LEARNING

For a single query, the high-level semantic information learned by long-term MSR and the low-level information learned by short-term learner are combined to improve the retrieval performance.

3.1. Short-term learner

The SVM active learning (SVM_{Active}) algorithm [7] is adopted as our short-term learner, whose output is a set of distance $\mathcal{D}^t = \{D^t(\mathbf{x})\}$. $D^t(\mathbf{x})$ denotes the distance of each image x to the decision boundary (x is "relevant" if $D^t(\mathbf{x}) > 0$, "irrelevant" otherwise). During each feedback round t, the learner selects the images with smallest $|D^t(\mathbf{x})|$ as $\mathcal{L}^t$ for user to label (*label set*), and select the images with largest $D^t(\mathbf{x})$ as the retrieval result $\mathcal{P}^t$ (*return set*) for this round. Actually $D^t(\mathbf{x})$ can be converted to represent the relevant degree of x to the current hidden concept c_q as:

$$E_s(\mathbf{x} \in c_q) = \frac{1}{Z} \exp\left\{D^t(\mathbf{x})\right\} \quad (2)$$

where Z is the normalization factor for E_s over x.

3.2. Long-term facilitation

During each feedback round, we first judge the semantic status S_C and S_L by algorithm in Fig.4, then select $\mathcal{P}^t$ and $\mathcal{L}^t$ as follows.

- **Return set selection**

Since an image x may belong to different concepts, we define the probability of x belonging to $\mathcal{C}_i$ or $\overline{\mathcal{C}}_i$ by:

$$p(\mathbf{x} \in \mathcal{C}_i) = \frac{N_{\mathbf{x} \in \mathcal{C}_i}}{\sum_{j=1}^{N} N_{\mathbf{x} \in \mathcal{C}_j}}, \quad p\left(\mathbf{x} \in \overline{\mathcal{C}}_i\right) = \frac{N_{\mathbf{x} \in \overline{\mathcal{C}}_i}}{\sum_{j=1}^{N} N_{\mathbf{x} \in \overline{\mathcal{C}}_j}}$$

where $N_{\mathbf{x} \in \mathcal{C}_i}$ ($N_{\mathbf{x} \in \overline{\mathcal{C}}_i}$) is the number of x being labeled to be in $\mathcal{C}_i$ ($\overline{\mathcal{C}}_i$) previously. And the probability that the current c_q equals to an existing c_i is defined by:

$$p(c_q = c_i) = \begin{cases} \delta(i = S_C) & , 1 \le S_L \le M,\ 1 \le S_C \le N \\ \frac{1}{|\mathcal{R}|} \sum_{\mathbf{x} \in \mathcal{R}} p(\mathbf{x} \in \mathcal{C}_i), & S_L = 0, S_C = 0 \\ 0 & , \text{otherwise} \end{cases}$$

where $\delta(\cdot)$ is Dirac function. Thus the relevant degree that an image x satisfies c_q is:

$$E_l(\mathbf{x} \in c_q) = \frac{1}{Z'} \exp\left\{\frac{1}{N} \sum_{i=1}^{N} [p(\mathbf{x} \in \mathcal{C}_i) - p(\mathbf{x} \in \overline{\mathcal{C}}_i)]\ p(c_q = c_i)\right\} \quad (3)$$

Where Z' is the normalization factor for E_l over x.

Then the short-term and long-term information are merged to represent the relevant degree of each image x to c_q as:

$$E_f(\mathbf{x} \in c_q) = \log E_s(\mathbf{x} \in c_q) + \log E_l(\mathbf{x} \in c_q) \quad (4)$$

Images with largest $E_f(\mathbf{x} \in c_q)$ form $\mathcal{P}^t$.

- **Label set selection**

When $S_L = 0$ and $S_C = 0$, we select $\mathcal{L}^t$ from the images in the compatible concepts in compatible layers with c_q. This helps the system to recognize the semantic status of c_q. When the semantic status is determined, $\mathcal{L}^t$ is selected with original $\mathcal{D}^t$ by mechanism of SVM$_{Active}$.

When a query session ends, we do long-term MSR learning by algorithm in Fig.5. The post processing is carried when every 100 query sessions have been taken.

4. EXPERIMENTS

The database has 10,000 real world images from Corel CDs, which come from 50 semantic categories, 200 images for each category. The low-level features used are color coherence in HSV space, the first three color moments in LUV space, the directionality texture feature, totally 145 dimensions. To make the experimental results more representative, instead of the usually used simulated experiments by ground truth, we ask 10 different users to launch totally 1000 queries by our system to construct the MSR for the database. The users have no special training but are only told to do retrieval to find what they want without changing the query concept during one query session.

- **The built MSR**

The final MSR built after 1000 query sessions has 88 concepts in 4 layers, whose structure is shown in Fig.7, where

Authorized licensed use limited to: Columbia University. Downloaded on December 29, 2008 at 22:41 from IEEE Xplore. Restrictions apply.

the semantic keywords are added after experiments for better understanding. We can see that the system learns many concepts adaptive to the database, which are not contained in ground truth semantics and have real world meanings. This also indicates that the built MSR can be used for annotation (we can simply annotate the concepts extracted by MSR to annotate the images in the database). Furthermore, when more images are added into the database, the already learned MSR is scalable to new added data by just adding the new data into the database part which is never labeled.

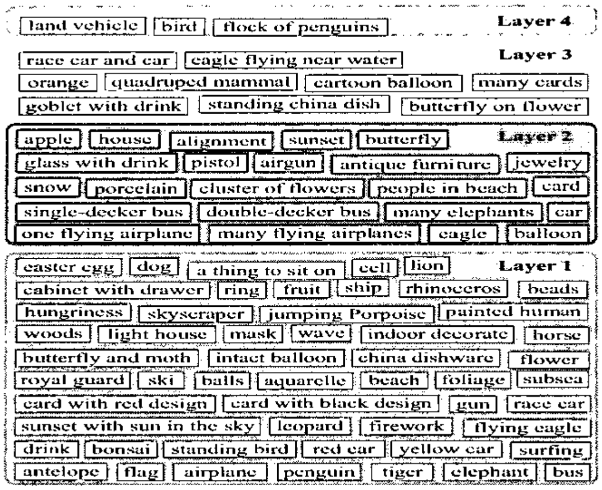


Fig. 7. MSR extracted after 1000 real retrieval sessions.

- **Precision evaluation**

With above built MSR, we use the ground truth semantics to test the performance improvement achieved by our algorithm over the 10,000 images. The top-k precision (P_k) (the percentage of "relevant" images in k size return set) is evaluated, and we carry totally 1000 independent queries to calculate the average result. Fig.8 gives the average P_{20} of our algorithm with different experience accumulation (after 100 to 1000 query sessions' learning), and the comparison with only short-term learning (original SVM$_{Active}$). The figure shows that the long-term MSR experience can significantly improve the retrieval performance, consistently from the first feedback round. For example the precision improvement for round 5 after 1000 queries is 28.11%. Also the advantage is more obvious as more queries are carried. Fig.9 gives a retrieval example querying for images belonging to "jumping porpoise" concept, which is not contained in ground truth concepts. The first screen of results after 3 feedback rounds for our method with final MSR accumulation and original SVM$_{Active}$ are listed. Where P_{20} for our method can attain 100%, for SVM$_{Active}$ is only 25%.

5. CONCLUSION

In this paper we have proposed a multi-layer semantic representation for database's hidden semantics and have implemented an algorithm to automatically build the MSR through long-term feedback process. Also, the MSR is incorporated with the short-term learner to significantly improve the retrieval result. Furthermore, the extracted semantics have real world meanings, which can be easily extended to further applications (such as facilitating semantic annotation).

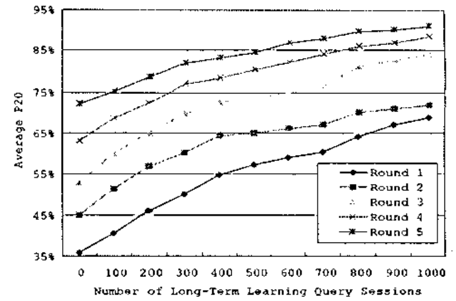
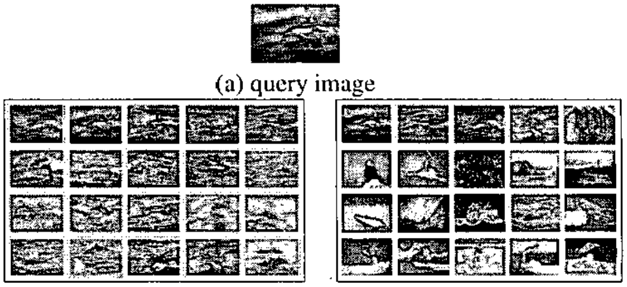


Fig. 8. Average retrieval P_{20} with and without MSR learning.



(a) query image

(b) $\mathcal{P}^3$ with MSR learning (c)$\mathcal{P}^3$ without MSR learning

Fig. 9. Retrieval example querying for "jumping porpoise", images with green frames are correct ones.

6. REFERENCES

[1] A.L. Dong and B. Bhanu, "A new semi-supervised EM algorithm for image retrieval," *IEEE Proc. of CVPR*, vol.2, pp.662-667, Madison, Wisconsin, 2003.

[2] A. Tvesky, "Feature of similarity," *Psychological review*, 84(4), pp. 327-352, 1977.

[3] C.S. Lee, W.Y. Ma and H.J. Zhang, "Information embedding based on user's relevance feedback for image retrieval," *Proc. SPIE of Multimedia Strage and Archiving Systems IV*, vol.3846, Boston, USA, 1999

[4] D.R. Heisterkamp, " Building a latent semantic index of an image database from patterns of relevance feedback," *IEEE Proc. of ICPR*, vol.4, 2002.

[5] J.W. Han. et al, "A memorization learning model for image retrieval," *Int. Conf. on Image Processing*, WP-S1, 2003.

[6] M.J. Li. et al, "A statistical correlation model for image retrieval," *3rd Intl Workshop on Multimedia Information Retrieval*, Ottwa, Canada, 2001.

[7] S. Tong, and E. Chang, "Support vector machine active learning for image retrieval," *ACM Multimedia*, Ottawa, Canada, 2001.

[8] Y. Rui. et al, "Relevance feedback: A powerful tool in interactive content-based image retrieval," *IEEE Trans. Circuit and System for Video Technology*, 8(5), pp.644-655, 1998.

Authorized licensed use limited to: Columbia University. Downloaded on December 29, 2008 at 22:41 from IEEE Xplore. Restrictions apply.