SAMPLING STRATEGIES FOR ACTIVE LEARNING IN PERSONAL PHOTO RETRIEVAL

Yi Wu, Igor Kozintsev, Jean-Yves Bouguet, Carole Dulong

Intel Corporation
2200 Mission College Blvd
Santa Clara, CA 95054, USA

ABSTRACT

With the advent and proliferation of digital cameras and computers, the number of digital photos created and stored by consumers has grown extremely large. This created increasing demand for image retrieval systems to ease interaction between consumers and personal media content. Active learning is a widely used user interaction model for retrieval systems, which learns the query concept by asking users to label a number of images at each iteration. In this paper, we study sampling strategies for active learning in personal photo retrieval. In order to reduce human annotation efforts in a content-based image retrieval setting, we propose using multiple sampling criteria for active learning: informativeness, diversity and representativeness. Our experimental results show that by combining multiple sampling criteria in active learning, the performance of personal photo retrieval system can be significantly improved.

1. INTRODUCTION

As the rate of personal digital media creation rises and storage becomes cheaper, the demand for solutions to manage personal photo collections is increasing tremendously. To this moment, existing image management systems mostly rely on keywords in a form of user annotation or the text that accompanies images for searching. Examples of such systems include web-based search provided by major search engine companies. This approach, however does not seem to be feasible for personal photo collections. The mere fact of targeting personal photos instead of catalog images brings several new challenges to the task of retrieval [1]. It is no longer reasonable to assume that a user would be willing or able to annotate consistently his or her complete image database. The tedious task of labeling appears even more difficult once we consider the very dynamic nature of a database of personal photographs. As a result, in this work, we are not considering query by keyword as a sole query modality for personal photo retrieval systems. Instead, we focus our attention on content-based queries where the only information that is readily available in the image file is used for search and retrieval. Such information used to compare images may consist of multiple visual features such as color, texture, objects, points of interest, etc.

A well-known problem of dealing with media data is the so-called "semantic gap". Once we choose not to rely on text information, we are faced with the problem of communicating user requests expressed in terms of high-level concepts (like "birthday party", "vacation trip") to a retrieval system that only "understands" language of low-level image features (e.g., color histogram, transform coefficients, etc.). To complicate things further, in personal photo databases the nature of stored data as well as query concepts for retrieval are highly user specific and often vary over time. Active learning-based image retrieval systems have recently received a lot of interest from academic and industry [2] because they offer a promising solution to the semantic gap problem and provide mechanism for online adaptation. Starting with one or several query examples, the interactive process allows the user to refine his/her request by giving feedback. Usually, the feedback from users to the systems is presented by binary labels indicating whether or not the photo belongs to the desired query.

In order to be practical, the active learning-based retrieval system cannot ask users to wait many iterations to get satisfactory results. Ideally, the feedback of the system should provide most useful (for the system) samples to the users to label and converge fast to the desired query concept. In this paper, our goal is to build a sampling strategy for active learning in the relevance feedback setting and efficiently retrieve relevant photos. To accelerate the learning of query concepts, we propose using multiple criteria: *informativeness, diversity and representativeness* that employed by the system to select the photos presented to the user at every feedback round. Our results for a realistic retrieval system and two personal photo databases of several thousand pictures demonstrate that this multi-criteria sampling mechanism can improve retrieval performance as compared to the state-of-the-art solutions published in literature.

2. MULTI-CRITERIA SAMPLING STRATEGIES FOR ACTIVE LEARNING

Active learning processes training data incrementally, using the model learned "so far" to select particularly useful exam-

1-4244-0367-7/06/$20.00 ©2006 IEEE

ples for labeling. Eventually, active learning methods reduce the number of instances that must be labeled to achieve a particular level of accuracy. Consider the problem of learning a binary classifier on a partially labeled database D. Let L be the labeled set and U be the unlabeled set ($U = D \setminus L$). The active learning system comprises of two stages:

- *A learning engine* to train a classifier on L. A learning engine is crutial for achieving good classification performance with limited training data.
- *A sampling engine* to select samples from U for users to label before passing it to the learning engine for next feedback iteration. A sampling engine is crutical for choosing the most valuable samples for users to label and converging to satisfied results fast.

Recent work on active learning often uses support vector machines (SVMs) [3] as the learning engine. In our system, we also employ SVMs as our learning engine because of their effectiveness in handling a small training data set. Given a set of training data with labels $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, where x_i is the i^{th} training instance and y_i is its class label (-1 denotes irrelevant and $+1$ denotes relevant), SVMs separate these two classes by a hyperplane with the maximum margin [3]. For nonlinearly separable cases, SVMs can project the training data onto a higher dimensional feature space via a Mercer kernel operator K. We can write $K(u, v) = \Phi(u) \cdot \Phi(v)$, where Φ is an input-to-feature space mapping function, and $\cdot$ denotes an inner product. The class prediction function for a data instance x is formulated as

$$f(x) = \sum_{i=1}^{p} \alpha_i y_i K(x, x_i) + b, \tag{1}$$

where α_i is the Lagrange multiplier of x_i and b is the offset.

The rest of the paper focuses on the sampling engine. Our objective is to control the labeling effort and accelerate the learning process by providing users the most valuable samples to label. In order to achieve the objective, we propose our sampling strategy based on three criteria: informativeness, diversity and representativeness.

2.1. Informativeness

This sampling strategy aims at selecting unlabeled data that can add most information to the current classifier. Tong and Chang proposed informativeness-based selection criterion in [4]. The basic idea is to select the most informative candidates whose representations in the feature space induced by the kernel are closest to the SVM hyperplane. In the other words, the data that have the prediction value $|f(x)|$ close to 0 are the most uncertain and informative samples. Given a set of unlabeled data $U = \{u_1, \cdots, u_n\}$, the informativeness of u_i is defined as

$$informativeness(u_i) = 1 - |f(u_i)|, \tag{2}$$

where the distance $|f(u_i)|$ has been normalized.

2.2. Diversity

Selecting examples exclusively based on the distances to the classification hyperplane might result high redundancy in the selected training set. Brinker [5] incoporated another sampling criteria of diversity, which encourages the selection of unlabeled samples that are far from the selected set and removes the redundancy within the selected samples. The redundancy of samples is measured by the angles between the samples.

Given a set of unlabeled data $U = \{u_1, \cdots, u_n\}$, the algorithm incrementally adds example u_i to the selected set S for labeling in next iteration. The diversity of u_i is defined as minimizing the redundancy between u_i and S:

$$diversity(u_i) = 1 - max_{s_j \in S} \frac{K(u_i, s_j)}{\sqrt{K(u_i, u_i) K(s_j, s_j)}}. \tag{3}$$

2.3. Representativeness

In addition to the most informative and diverse examples, we also prefer the most representative examples from the unlabeled pool. The examples with high representativeness will add more information to the training set. The representativeness of an example can be evaluated on how many examples are similar to it. Given a set of unlabeled data $U = \{u_1, \cdots, u_n\}$, the representativeness of u_i is defined as the average similarity between u_i and all the other data in U.

$$representativeness(u_i) = \frac{\sum_{j \neq i} K(u_i, u_j)}{n - 1} \tag{4}$$

2.4. Multi-criteria Sampling

In order to combine these three criteria and strike a proper balance between them, we propose multi-criteria sampling strategy. We incrementally construct a new training batch S from the unlabeled data pool as Figure 1 shows. We combine the informativeness, diversity and representativeness criteria using the function $w_1 \times informativeness(x_i) + w_2 \times diversity(x_i) + w_3 \times representativeness(x_i)$. The individual importance of each criterion is adjusted by w_1, w_2 and w_3. We add the candidate example x_i to the selected labeling data set S one by one until the size of S grows to the predefined value h.

3. EFFICIENT IMPLEMENTATION

To select new examples, a naive way to calculate the *diversity* value of each candidate is to evaluate it against all the data already added in S, which results in a quadratic dependence of computational time on h, the size of the new selected batch.

```
Input:
$U = \{u_1, \cdots, u_n\}$; /* A set of unlabeled data */
$h$; /* the size of selected data for labeling */
$w_1, w_2, w_3$; /* weights of three criteria */
Output:
$S$; /* A set of selected data for labeling */
Function calls:
$representativeness(u_i)$; /* representativeness of $u_i$ */
$diversity(u_i)$; /* redundancy of $u_i$ */
$informativeness(u_i)$; /* informativeness of $u_i$ */

Begin:
1)  $S = \emptyset$; /* Initialization */
2)  repeat
3)      $selected = argmax_{i \in U \setminus S}(w_1 \times informativeness(u_i) + w_2 \times diversity(u_i) + w_3 \times representativeness(u_i))$;
4)      $S = S \bigcup \{u_{selected}\}$;
5)  until card($S$)=$h$;
6)  return $S$;
End
```

Fig. 1. Sample Selection Strategy.

An efficient implementation of *diversity* evaluation was proposed in [5]. The main idea is to cache the diversity values for all $card(U \setminus S)$ unlabeled examples and perform an update if the cosine of the angle between an unlabelled example and a newly added example is greater than the stored maximum.

To calculate the *representativeness* value of each candidate, a naive way is to evaluate it against all $card(U \setminus S)$ unlabeled examples, which results in $O(hn^2)$ computational cost in each feedback iteration (n is the number of unlabeled data.) It is more efficient to cache the *representativeness* values for all $card(U \setminus S)$ unlabeled examples and perform an update when a new example is added to S.

The complete pseudo code of an efficient implementation of the sample selection strategy is given in Figure 2. Before any feedback, the representativeness of each data is calculated using Equation 4 and saved in the array of $repre$. In each feedback iteration, this array will be passed to the algorithm as inputs and updated when new data is added to the labeling pool (step 15).

```
Input:
$U = \{u_1, \cdots, u_n\}$; /* A set of unlabeled data */
$repre$; /* An array of representativeness values for unlabeled data */
$h$; /* the size of selected data for labeling */
$w_1, w_2, w_3$; /* weights of three criteria */
Output:
$S$; /* A set of selected data for labeling */
Variable:
$info = array[n]$ of double;
$maxCos = array[n]$ of double;
Function calls:
$swap(i, j)$; /* swap all associated values at position i and j */
$informativeness(u_i)$; /* informativeness of $u_i$ */
$K(u_i, u_j)$; /* kernel value between $u_i$ and $u_j$ */

Begin:
/* Initialization */
1)  $S = \emptyset$;
2)  for $i = 1$ to $n$ do
3)      $info[i]$ = $informativeness(u_i)$;
4)      $maxCos[i]$=0;
5)  end for

/* select examples for labeling until card(S)=h */
6)  for $k = 1$ to $h$ do
7)      $selected = argmax_{i \in U \setminus S}(w_1 \times info[i] + w_2 \times (1 - maxCos[i]) + w_3 \times repre[i])$;
8)      $S = S \bigcup \{u_{selected}\}$;

/* swap the selected example and the $k^{th}$ unlabeled example */
9)      swap($k$, $selected$);

/* update diversity value and representativeness value */
10)     for $j = k + 1$ to $n$ do
11)         cos = $\frac{K(u_k, u_j)}{\sqrt{K(u_k, u_k)K(u_j, u_j)}}$;
12)         if cos $> maxCos[j]$ then
13)             $maxCos[j]$= cos;
14)         end if
15)         $repre[j] = (repre[j] \times (n - k) - K(u_k, u_j))/(n - k - 1)$;
16)     end for

17) end for
18) return $S$;
End
```

Fig. 2. An Efficient Implementation of Sample Selection.

4. EXPERIMENTAL RESULTS

In this section, we will evaluate the effectiveness of our proposed multi-criteria sampling strategy for active learning in personal photo retrieval. We conducted our experiments on two personal photo datasets contributed from our researchers. The first dataset, named as **DI**, contains around $5k$ photos taken over five years. The second dataset, named as **DII**, contains around $2k$ photos taken over three months. The owner of **DI** proposed 21 queries she was interested in. The owner of **DII** proposed 22 queries that he was interested in. The query semantics varies from *object* and *place* queries to complex *event* queries. The percentage of photos in the dataset relevant to each query varies from 0.03% to 2%.

We utilized Scale-invariant feature transform (SIFT) [6] descriptors as feature representation. Each photo is described by a set of feature points, and each feature point is described by a 128 dimensional feature vector. For measuring similarity between photos, we have designed a similarity metric based on feature point match in the same flavor of the work from Grauman and Darrell [7].

For each query session, we randomly selected one relevant photo and 19 irrelevant photos as the initial training batch (RF=0). The system presented photos in decreasing order of relevance. At the same time, the system returned 20 images for users to label. Users were expected to label them as relevant or irrelevant to the query. As labelled data added to the training pool, the classifier were re-trained and generated a new list of relevant photos. This relevance feedback were repeated for five iterations (RF=1,...,5). The performance was

ITERATIONS	I	$I+D$	$I+R$	$D+R$	$I+D+R$
RF=0	0.295				
RF=1	0.419	0.418	0.456	0.306	0.462
RF=2	0.528	0.533	0.549	0.324	0.557
RF=3	0.582	0.582	0.576	0.332	0.605
RF=4	0.611	0.624	0.625	0.341	0.661
RF=5	0.640	0.646	0.654	0.344	0.682

Table 1. MAPs for 21 queries of **DI**.

ITERATIONS	I	$I+D$	$I+R$	$D+R$	$I+D+R$
RF=0	0.345				
RF=1	0.527	0.538	0.538	0.350	0.568
RF=2	0.620	0.623	0.626	0.363	0.646
RF=3	0.678	0.690	0.699	0.377	0.702
RF=4	0.714	0.736	0.732	0.400	0.751
RF=5	0.743	0.762	0.766	0.412	0.785

Table 2. MAPs for 22 queries of **DII**.

evaluated based on the sorted photo list after each iteration and averaged over 10 runs. The retrieval system performance is measured using NIST average precision (AP), which gives a global evaluation of the system over all the precision and recall values [8].

Table 1 shows the mean average precision (MAP) for 21 queries of **DI** at each feedback iteration. We evaluated five different sampling strategies for selecting data to label: using *informativeness* criterion only (I), using *informativeness* and *diversity* criteria ($I + D$), using *informativeness* and *representativeness* criteria ($I + R$), using *diversity* and *representativeness* criteria ($D + R$), and using *informativeness, diversity* and *representativeness* criteria ($I + D + R$). In our experiments, the individual importance of each criterion *informativeness, diversity* and *representativeness* was set as $w_1 = 0.47$, $w_2 = 0.20$ and $w_3 = 0.33$.

For all these five methods, the MAPs for the initial retrieval (RF=0) were the same because there was no feedback at that stage. When only using *diversity* and *representativeness* criteria ($D + R$), the performance didn't get too much improvement by doing relevance feedback. And the overall performance was the worst among all these five approaches. The possible reason is that these two criteria mainly focus on optimizing data distribution of the selected training data, but might not add more information to the classifier. For the other four methods, the performance of the retrieval system improved tremendously during relevance feedback.

Compared to using *informativeness* criteria only (I), both $I+D$ and $I+R$ could perform better. However, the most significant improvement was observed when all the three criteria ($I + D + R$) were combined.

Table 2 shows the mean average precision (MAP) for 22 queries of **DII** at each feedback iteration. We also compared the results when employing five sampling strategies: I, $I+D$, $I + R$, $D + R$, and $I + D + R$. Similar to **DI**, when only using *diversity* and *representativeness* criteria ($D + R$), the performance was the worst among all these five approaches. When all the three criteria ($I + D + R$) were combined, the performance was the best.

The experimental results show that a) *informativeness* criterion is the most important criteria by adding useful information to the classifier; b) *diversity* and *representativeness* criteria can optimize data distribution and reduce the redundancy of the selected training data; c) combining *informativeness, diversity* and *representativeness* criteria as the sampling strategy for active learning will help learn the query concept more quickly and cost less in annotation.

5. CONCLUSIONS

In this paper, we have studied the sampling strategies for active leraning in personal photo retrieval. Our algorithm selects samples to label in each feedback iteration by combining three criteria *informative, diverse* and *representative*. This approach provides a fast method to obtain better photo retrieval performance and costs less human labeling effort. Preliminary experiments show that by considering multiple sampling criteria in active learning, the performance of photo retrieval system can be improved. In our future work, we will experiment our active learning sampling strategy with other learning engines besides support vector machines.

6. REFERENCES

[1] J-Y. Bouguet, C. Dulong, I. Kozintsev, and Y. Wu, "Requirements for benchmarking personal image retrieval systems," *Proceedings of the SPIE/IST Conference on Internet Imaging VII*, January 2006.

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

[3] J. C. Burges Christopher, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, 1998.

[4] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," *Proc. of the ninth ACM international conference on Multimedia*, 2001.

[5] K. Brinker, "Incorporating diversity in active learning with support vector machines," *Proceedings of the Twentieth International Conference on Machine Learning*, pp. 59–66, 2003.

[6] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 2, no. 60, pp. 91–110, 2004.

[7] K. Grauman and T. Darrell, "The pyramid match kernel: Discriminative classification with sets of image features," *Proceedings of the IEEE International Conference on Computer Vision, Beijing, China*, October 2005.

[8] NIST, "Common evaluation measures," 2001.