Autonomous Visual Model Building based on Image Crawling through Internet Search Engines

Xiaodan Song
Department of Electrical Engineering
University of Washington, Box 352500, Seattle, WA 98195, USA
song@ee.washington.edu

Ching-Yung Lin
IBM T.J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532, USA
cylin@watson.ibm.com

Ming-Ting Sun
Department of Electrical Engineering
University of Washington, Box 352500, Seattle, WA 98195, USA
sun@ee.washington.edu

ABSTRACT
In this paper, we propose an autonomous learning scheme to automatically build visual semantic concept models from the output data of Internet search engines without any manual labeling work. First of all, images are gathered by crawling through the Internet using a search engine such as Google. Then, we model the search results as "Quasi-Positive Bags" in the Multiple-Instance Learning (MIL) framework. We call this generalized MIL (GMIL). We propose an algorithm called "Bag K-Means" to find the maximum Diverse Density (DD) without the existence of negative bags. A cost function is found as K-Means with special "Bag Distance". We also propose a solution called "Uncertain Labeling Density" (ULD) which describes the target density distribution of instances in the case of quasi-positive bags. A "Bag Fuzzy K-Means" is presented to get the maximum of ULD. By this generalized MIL with ULD, the model for a particular concept is learned from the crawled images of the Internet search engines. Experiments show that our algorithm can get correct models for the concepts we are interested in. Compared to the original Google Image Search, our algorithm shows improved accuracy.

Categories and Subject Descriptors
H.3.3 [**Information Storage And Retrieval**]: Information Search and Retrieval—*retrieval models*

General Terms
Algorithms

Keywords
Content-based Image Retrieval, Cross-Modality, Automatic Training, Multiple-Instance Learning, Uncertain Labeling Density, Quasi-Positive Bag, Image Crawling

1. INTRODUCTION
As the amount of image data increases, content-based image indexing and retrieval is becoming increasingly important. Semantic model-based indexing has been proposed as an efficient method, which matches human experience in search. Supervised learning has been used as a successful method to build generic semantic models [11]. This approach performed best in the NIST TRECVID concept detection benchmarking in 2002 and 2003 [17][11]. However, in this approach, tedious manual labeling is needed to build tens or hundreds of models for various visual concepts. For example, in 2003, 111 researchers from 23 institutes spent 220+ hours to annotate 63 hours of TREC 2003 development corpus [16]. This manual annotating process is time- and cost- consuming, and, thus, makes the system hard to scale. Even with this enormous labeling effort, any new instances not previously labeled would not be able to be dealt with. Semi-supervised learning or partial annotation was proposed to reduce the involved manual effort [21][22]. Once the database is partially annotated, traditional pattern classification methods are often used to derive semantics of the objects not yet annotated. However, it is not clear how much annotation is sufficient for a specific database, and what the best subset of the objects to be annotated is. It is desirable to have an automatic learning algorithm, which totally does not need the costly manual labeling process.

The work we proposed in [1] tries to solve this problem by making use of the correlation between audio and visual data in video sequences. The correlation between the textual and the visual modalities for the huge amount of image data available on the web would be another possibility for our autonomous learning scheme to build models for concepts for content-based retrieval. Recently, some Internet search engines have supported image searches. Among them, Google's Image Search is the most comprehensive on the Web, with more than 150 million images indexed and available for viewing. Google gathers a large collection of images for its search engine by analyzing the text on the page adjacent to the image, the image caption, and dozens of other factors to determine the image content. Google also uses sophisticated algorithms to remove duplicates, and to ensure that the most relevant images are presented first in the results. Traditionally, relevance feedback technique is involved for image retrieval based on these imperfect data [18][19][20]. Widely used in text retrieval [23][24], relevance feedback was first proposed by Rui et al [18] as an interactive tool in content-based image retrieval. Then, it becomes a powerful tool and a major focus of research for bridging the gap between low-level features and high-level semantics. During the retrieval process, the user interactively selects the most relevant images and provides a weight according to the preference for each relevant image. By

MIR'04, October 15–16, 2004, New York, New York, USA.

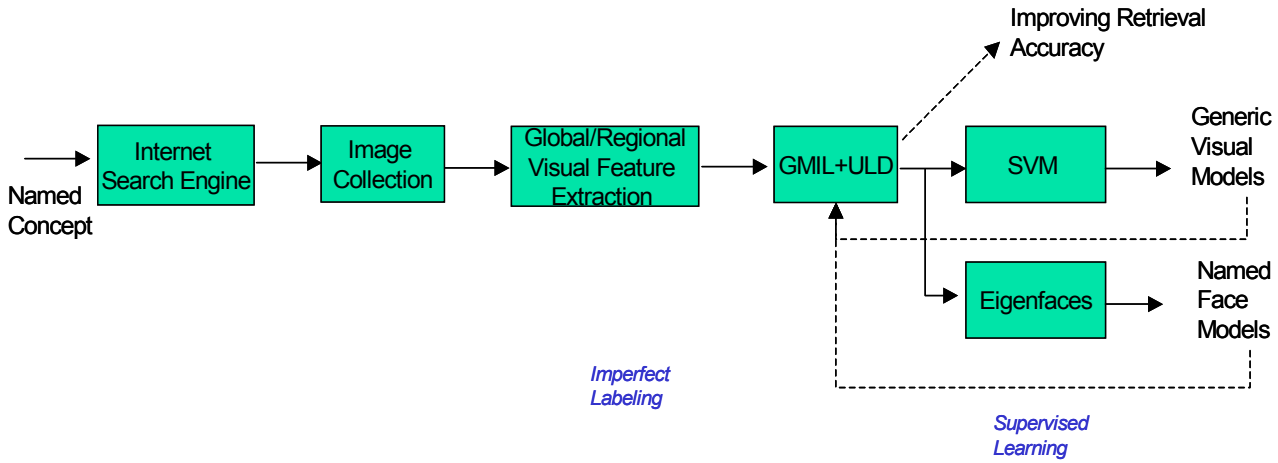


Figure 1. The framework for autonomous concept learning based on image crawling through Internet search engines

dynamically updating weights based on the feedback, user's high level query and perception subjectivity is captured. Relevance feedback moves the query point towards the relevant objects or selectively weighs the features in the low-level feature space based on user feedback. However, relevance feedback still needs human involvements. Thus, it is very difficulty, if not impossible, to build a large amount of models based on relevance feedback. In this paper, we propose a solution to this major obstacle of machine learning. We show that it is possible to automatically build up the models without any human intervention for various concepts for future search and retrieval tasks.

Our scheme is based on the Multiple Instance Learning (MIL) approach. MIL was recently proposed for machine learning to solve the ambiguity in the manual labeling process by making weaker assumptions about the labeling information [2][3][4]. In this learning scheme, instead of giving the learner labels for individual examples, the trainer only labels collections of examples, which are called bags. A bag is labeled negative if all the examples in it are negative. It is labeled positive if there is at least one positive example in it. The key challenge in MIL is to cope with the ambiguity of not knowing which instances in a positive bag are actually positive and which are not. Based on that, the learner attempts to find the desired concept.

MIL helps to deal with the ambiguity in the manual labeling process. However, users still have to label the bags in the MIL framework. To prevent the tedious manual labeling work, we need to generate the positive bags and negative bags automatically. In practical applications, it is very difficult if not impossible to generate the positive bags reliably. Also, negative bags are often not available. In this paper, we propose a Generalized MIL (GMIL) concept by introducing "Quasi-Positive bags" to remove the strong requirement of using strictly positive bags in the MIL framework. In the GMIL framework, we also avoid the strong dependency on the negative bags. Maron et al. proposed a Diverse Density algorithm as an efficient solution for MIL [2]. In this paper, we first propose an efficient algorithm called "Bag K-Means" to find the maximum Diverse Density (DD) with the absence of negative bags. We develop a cost function, which uses K-Means with a special "Bag Distance". We also propose "Uncertain Labeling Density" (ULD) to resolve the "quasi-positive bags" issues in the generalized MIL problem. Compared to DD, ULD pays more attention to the structure of the "Quasi-Positive bags" instead of depending on the distribution of the negative instances like many traditional MIL algorithms do. A "Bag Fuzzy K-Means" is proposed to efficiently get the maximum of ULD. Compared to what we proposed in [1], a more general formulation for ULD and theoretical analysis is given in this paper. Based on our proposed GMIL and ULD approach, we propose an automatic learning scheme to generate models automatically for various concepts from cross-textual and visual information.

The overall process of the cross-modality automatic learning scheme is shown in Figure. 1. First of all, images are gathered by image crawling from the Google search results. Then, using the GMIL solved by ULD, the most informative examples are learned and the model of the named concept is built. This learned model can be used for concept indexing in other test sets. One of the applications is to use it as a "quasi-relevance feedback" mechanism, which can be used to improve the accuracy of the original retrieved image dataset. For instance, a revised relevance score rank list can be generated by the distance from the model and the retrieved image dataset. This can also be used to improve the retrieval accuracy.

The rest of this paper is organized as follows. In Section 2, we briefly review MIL and generalize it by introducing "Quasi-Positive bags" so that the learning process can be done based on the cross-modality correlation without any manually labeling work. In Section 3, DD for solving the MIL problem is introduced. The MIL is then generalized to allow false-positive bags, and ULD is proposed to solve the generalized MIL problem. Both theoretical and experimental analyses are given for ULD. The details of our autonomous learning algorithm are described in Section 4. Finally, experimental results and conclusions are given in Sections 5 and 6 respectively.

2. GENERALIZED MULTIPLE-INSTANCE LEARNING

In this section, we present a brief introduction to Multiple-Instance Learning, and generalize it for autonomous learning by introducing the concept of "Quasi-Positive Bags".

2.1 Multiple-Instance Learning

Given a set of instances $x_1, x_2 ..., x_N$, the task in a typical machine learning problem is to learn a function

$$y = f(x_1, x_2, \cdots, x_N) \quad (1)$$

so that the function can be used to classify the data. In traditional supervised learning, training data are given in terms of (y_i, x_i) to learn the function for classifying the data outside the training set. In MIL, the training data are grouped into bags $X_1, X_2, \cdots, X_M$, with $X_j = \{\mathbf{x}_i : i \in I_j\}$ and $I_j \subseteq \{1, \cdots K\}$. Instead of giving the labels y_i for each instance, we have the label Y_i for each bag. A bag is labeled negative ($Y_j = -1$), if all the instances in it are negative. A bag is positive ($Y_j = 1$), if at least one instance in it is positive.

The MIL model was first formalized by Dietterich et al. [5] to deal with the drug activity prediction problem. Following that, an algorithm called Diverse Density (DD) was developed in [3] to provide a solution to MIL, which performs well on a variety of problems such as drug activity prediction, stock selection, and image retrieval [4]. Later, the method is extended in [6] to deal with the real-valued labels instead of the binary labels. Many other algorithms, such as k-*NN* algorithms [7], Support Vector Machine (SVM) [8], and EM combined with DD [15] are proposed to solve MIL. However, most of the algorithms are sensitive to the distribution of the instances in the positive bags, and cannot work without negative bags.

In the MIL framework, users still have to label the bags. To prevent the tedious manual labeling work, we need to generate the positive bags and negative bags automatically. However, in practical applications, it is very difficult if not impossible to generate the positive and negative bags reliably. Without reliable positive and negative bags, DD may not give reliable solutions. To solve the problem, we generalize the concept of "Positive bags" to "Quasi-Positive bags", and propose "Uncertain Labeling Density" (ULD) to solve this generalized MIL problem.

2.2 Quasi-Positive Bag

In our scenario, although there is a relatively high probability that the concept of interest (e.g. a person's face) will appear in the crawled images, there are many cases that no such association exists (e.g. Figure. 3 in Section 5). If these images are used as the positive bags, we may have false-positive bags that do not contain the concept of interest. In this case, DD may not be able to give correct results as will be shown later. To overcome this problem, we extend the concept of "Positive bags" to "Quasi-Positive bags". A "Quasi-Positive bag" has a high probability to contain a positive instance, but may not be guaranteed to contain one. The introduction of "Quasi-Positive bags" removes a major limitation of applying MIL to many practical problems.

Definition: Generalized Multiple Instance Learning (GMIL)

In the generalized MIL, a bag is labeled negative ($Y_j = -1$), if all the instances in it are negative. A bag is Quasi-Positive ($Y_j = 1$), if in a high probability, at least one instance in it is positive.

3. DIVERSE DENSITY and UNCERTAIN LABELING DENSITY

In this section, we first give a brief overview of Diverse Density proposed by Moron et al. [2]. We show that it has a similar cost function as the K-Means algorithm but with a different definition of distance, which we call "bag distance". Then, an efficient Bag K-Means algorithm is presented to efficiently find the maximum of DD instead of using the time-consuming gradient descent algorithm. We also prove the convergence property of this Bag K-Means algorithm. This algorithm can be used to find the maximum DD solutions in MIL with the existence of positive bags but without the negative bags. Then, for the GMIL, we introduce a concept called Uncertain Labeling Density (ULD) to solve the problem of quasi-positive bags. A Bag Fuzzy K-Means algorithm is presented to find the maximum of ULD.

3.1 Diverse Density

One way to solve MIL problems is to examine the distribution of the instance vectors, and look for a feature vector that is close to the instances in different positive bags and far from all the instances in the negative bags. Such a vector represents the concept we are trying to learn. This is the basic idea of the Diverse Density algorithm [2].

Diverse Density is a measure of the intersection of the positive bags minus the union of the negative bags. By maximizing Diverse Density, we can find the point of intersection (the desired concept). Here a simple probabilistic measure of Diverse Density is explained. We use the same notations as in [2]. We denote *i*th positive bag as B_i^+, the *j*th instance in that bag as B_{ij}^+, and the *j*th instance from a negative bag as B_i^-. Assume the intersection of all positive bags minus the union of all negative bags is a single point *t*, we can find this point by

$$\arg\max_t \prod_i \Pr(t \mid B_i^+) \prod_i \left(1 - \Pr(t \mid B_i^-)\right). \quad (2)$$

This is the formal definition of Diverse Density. $\Pr(t \mid B_i)$ is estimated by the most-likely-cause estimator, in which only the instance in the bag which is most likely to be in the concept c_t is considered:

$$\Pr(t \mid B_i) = \max_j \left\{\Pr(t \mid B_{ij})\right\} \quad (3)$$

The distribution is estimated as a Gaussian-like distribution of:

$$\Pr(t \mid B_{ij}) = \exp\left(-\left\|B_{ij} - t\right\|^2\right) \quad (4)$$

where $\left\|B_{ij} - t\right\|^2 = \sum_k \left(B_{ijk} - t_k\right)^2$. For the convenience of discussion, we define "Bag Distance" as:

$$d_i^t \equiv \min_j \left\|B_{ij} - t\right\|^2 \quad (5)$$

3.2 The Bag K-Means Algorithm for Diverse Density with the absence of negative bags

In our special application, where negative bags are not provided, (2) can be simplified as:

$$\arg\max_t \prod_i \Pr(t \mid B_i^+) = \arg\min_t \sum_i d_i^t \quad (6)$$

which is to minimize a metric of sum of the average distance to the centroid. It has exactly the same form of the cost function as K-Means' but with a different definition of d in (5). We call it Bag K-Means in this paper. Basically, when there is no negative bag, the DD algorithm is trying to find the centroid of the cluster by K-Means when $K=1$. With this conclusion, we propose an efficient algorithm to find the maximum DD by the Bag K-Means algorithm as follows:

(1) Choose an initial seed t

(2) Choose a convergence threshold ε

(3) For each bag i, choose one example s_i which is closest to the seed t, and calculate the distance d_i^t

(4) Calculate $t_{new} = \sum_i s_i \Big/ N$, where N is the total number of bags.

(5) If $\|t - t_{new}\| \le \varepsilon$, stop, otherwise, update $t = t_{new}$, and repeat (3) to (5).

The algorithm starts with an initial guess of the target point t which is obtained by trying instances from Qusi-Positive bags, then an interactive searching algorithm is performed to update the position of this target point t so that equation (6) is achieved.

We now provide the proof of convergence of Bag K-Means algorithm.

Theorem: The Bag K-Means algorithm converges.

Proof: Assume t_i is the centroid we found in the iteration i, and s_{ij} is the sample obtained in step (3) for bag j. By step (4), we get a new centroid t_{i+1}. We have:

$$\sum_j \|s_{ij} - t_{i+1}\|^2 \le \sum_j \|s_{ij} - t_i\|^2 \tag{7}$$

with the property of the traditional K-Means algorithm.

Because of the criterion of choosing new s_{i+1j}, we have:

$$\sum_j \|s_{i+1j} - t_{i+1}\|^2 \le \sum_j \|s_{ij} - t_{i+1}\|^2 \tag{8}$$

Combine (7) and (8), we get

$$\sum_j \|s_{i+1j} - t_{i+1}\|^2 \le \sum_j \|s_{ij} - t_i\|^2 , \tag{9}$$

which means the algorithm decreases the cost function J in (6) each time. Therefore, this process will converge.

3.3 Uncertain Labeling Density

In our generalized MIL, what we have are Quasi-Positive bags, i.e., some false-positive bags do not include positive instances at all. In a false-positive bag, by the original DD definition, $\Pr(t \mid B_i^+)$ will be very small or even zero. These outliers will influence the DD significantly due to the multiplication of the probabilities. Based on our previous deduction, which proves the equalization between the DD when there is no negative bag and the proposed Bag K-Means algorithm, this outlier problem is a correspondence of the challenging outlier problem to the traditional K-Means algorithm [9][10]. Many algorithms have been proposed to handle this outlier problem in K-Means. Among them, fuzzy K-Means algorithm is the most well known [9][10]. The intuition of the algorithm is to give different measurements (weights) on the relationship each example belonging to any cluster. The weights indicate the possibility a given example belongs to any cluster. By assigning low weight values to outliers, the effect of noisy data on the clustering process is reduced. In this paper, based on this similar idea from fuzzy K-Means, we propose an Uncertain Labeling Density (ULD) algorithm to handle the Quasi-Positive bag problem for GMIL.

Definition: Uncertain Labeling Density (ULD)

$$ULD(t) = \prod_i \left(\Pr\left(t \mid B_i^+\right)^{\left(\mu_i^t\right)^b} \right)$$
$$\sum_i \mu_i^t = N \tag{10}$$

where μ_i^t represents the weight of bag i belonging to concept t, and $b > 1$ is the fuzzy exponent. It determines the degree of fuzziness of the final solution. Usually, $b = 2$.

Similarly, we get the conclusion that the maximum of ULD can be obtained by Fuzzy K-Means with the definition of "Bag Distance" (5), with maximizing the cost function:

$$\arg\max_t \prod_i \Pr\left(t \mid B_i^+\right)^{\left(\mu_i^t\right)^b} = \arg\min_t \sum_i \left(\mu_i^t\right)^b d_i^t \tag{11}$$

3.4 The Bag Fuzzy K-Mean Algorithm for Uncertain Labeling Density

The Bag Fuzzy K-Means algorithm is proposed as follows:

(1) Choose an initial seed t among the Quasi-Positive bags

(2) Choose a convergence threshold ε

(3) For each bag i, choose one example s_i which is closest to t this seed, and calculate the Bag Distance d_i^t

(4) Calculate

$$t_{new} = \frac{\sum_{i=1}^{N} \left(\mu_i^t\right)^b s_i}{\sum_{i=1}^{N} \left(\mu_i^t\right)^b}$$
$$\mu_i^t = \frac{N\left(1/d_i^t\right)^{1/(b-1)}}{\sum_{j=1}^{N} \left(1/d_j^t\right)^{1/(b-1)}} \tag{12}$$

where N is the total number of bags.

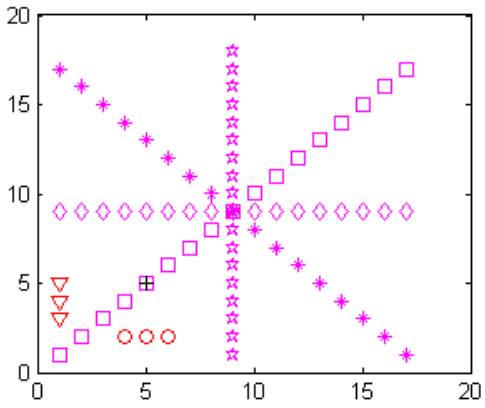
NOTE: In practice, we add a small number ε' to d_i^t to avoid the situation of divided by 0.

(5) If $\|t - t_{new}\| \le \varepsilon$, stop, otherwise, update $t = t_{new}$, and repeat (3) to (5).

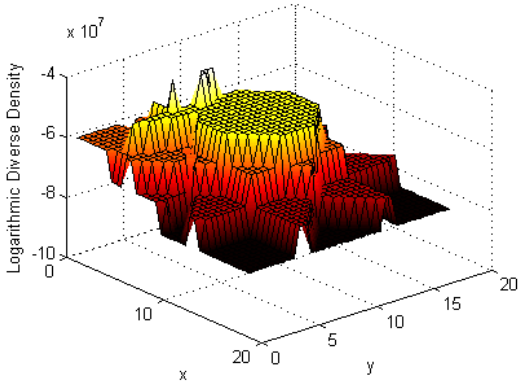
Essentially, the weights indicate the possibility an instance belongs to the interested cluster. By assigning low weights to outliers, the effect of them on the clustering process is reduced. In each step, the weight of each instance is updated according to the distance to the centroid t. And the updated weighted mean is set as the current centroid. The convergence of this Bag Fuzzy K-Mean algorithm can be obtained by the previous proof of the Bag K-Means algorithm and the convergence of the original Fuzzy K-Means algorithm.

Figure 2 shows an example with Quasi-Positive bags, and without negative bags. Different symbols represent various Quasi-Positive bags. There are two false-positive bags, which are illustrated by the inverse-triangles and circles, in this example. The true intersection point is the instance with the

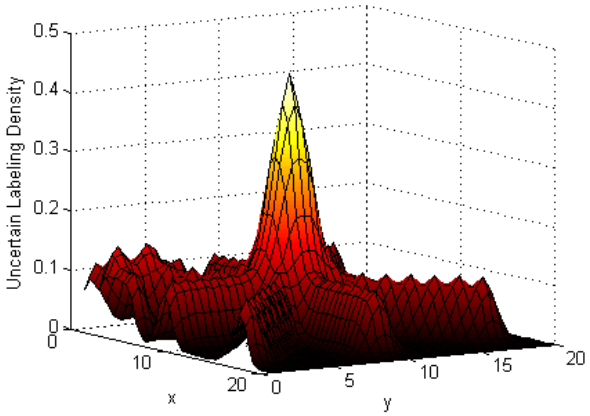
value (9, 9) with intersections from four different positive bags. Just by finding the maximum of the original Diverse Density, the algorithm will converge to (5, 5) (labeled with a "+" symbol) because of the influence of the false-positive bags. Figure 2(b) illustrates the corresponding Diverse Density values. By using the ULD method, it is easy to obtain the correct intersection point with the ULD as showing in Figure. 2(c).



(a) An example with Quasi-Positive bags



(b) The corresponding Diverse Density



(c) Using Uncertain Labeling Density

Figure 2. Comparison of MIL using Diversity Density and Uncertain Labeling Density Algorithms in the case of quasi-positive bags

4. CROSS-MODALITY AUTOMATIC TRAINING

In this section, we describe the features we used and how to automatically generate the quasi-positive bags in our scheme. In this paper, we only show the procedure of the cross-modality training on face models. For generic visual models, the system can use a region segmentation, feature extraction and supervised learning framework as in [17].

4. 1 Feature Generation

We focus on the frontal face model. We first extract frontal faces from the images obtained from the search engine, use skin detection to exclude some false alarm detections, and then obtain the projection coefficients based on eigenfaces for the face recognition.

4.1.1 Face detection

The face detection algorithm we used is based on the approach proposed in [12], which extends Viola et al.'s rapid object detection scheme [13]. It is based on a boosted cascade of simple features by enriching the basic set of simple Haar-like features and incorporating a post optimization procedure. This algorithm reduces the false alarm rate significantly with a relative high hit-rate and fast speed. However, there are still some false detections since it is based on gray value features only. We propose to reduce those false alarms by skin color detection. Our skin detection algorithm is based on a skin pixel classifier, which is derived using the standard likelihood ratio approach in [14]. After getting skin pixel candidates, we post-process the candidates to determine the skin regions, using techniques including Gaussian blurring, thresholding, and mathematical morphological operations such as closing and opening.

4.1.2 Eigenface generation

The eigenfaces we use in this paper is the same as what we obtained in [1]. The frontal faces, which are in a relatively large scale (larger than 48×48) and include certain skin regions (face regions which cover more than a quarter of the whole image), are detected from the crawled images. After normalized to a size of 64×64 and a median value 128 of gray level, they are used to get the top 22 eigenfaces with 85% energy for recognition. The features used throughout this paper are the projection coefficients based on these eigenfaces.

4.2 Quasi-positive bag generation

The quasi-positive bags are those gathered images with the extracted frontal faces as the instances. An illustration of the quasi-positive bags is shown in the bottom part of Figure 3.

5. EXPERIMENTAL RESULTS

We applied our algorithm to learn models of four particular persons, Bill Clinton, Hillary Clinton, Newt Gingrich, and Madeleine Albright. Figure 3 shows the dataflow in our scheme. First of all, a name is typed in Google Image Search Engine, such as "Bill Clinton". Then, an image crawler is applied to the resultant images from the search. These images were gathered in May 2004. The gathered images are in the form of .jpg or .gif.

Because most .gif images are just animations, we only consider jpeg images as the experimental data. After that, faces are extracted from those images automatically and the faces from the same image constitute a Quasi-Positive bag. Then, the most informative example for that person is learned by our proposed GMIL by ULD algorithm and a rank list is generated based on the distance from this example.

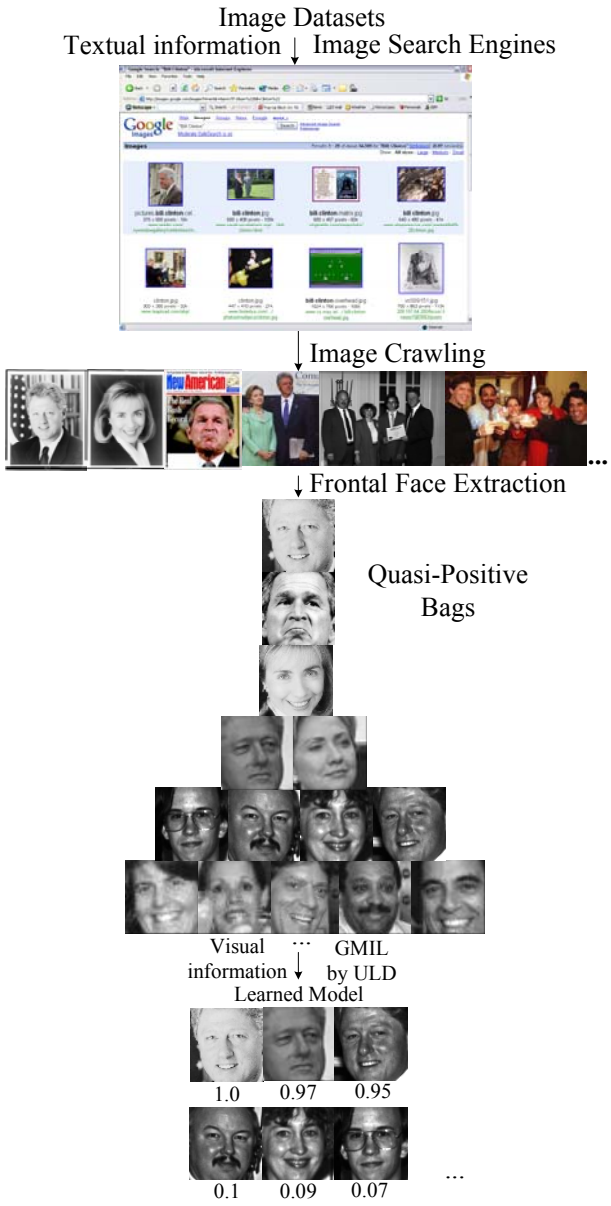
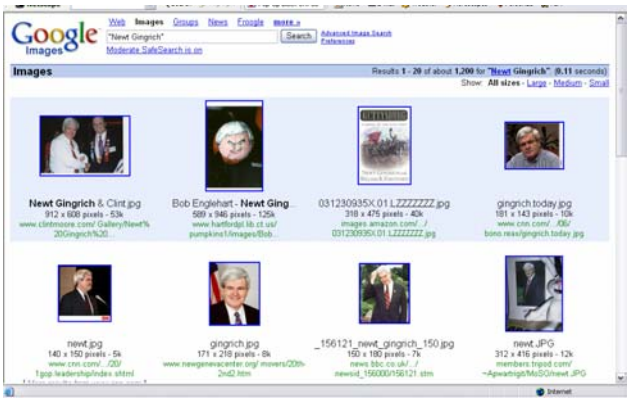


Figure 3. An example of building the face model of "Bill Clinton"

To illustrate the performance of our algorithm, we compare the top ranked retrieved images as well as the precision values. Figure 4 and Figure 5 show part of the original top 8 searched results by Google, and the top 10 images in the rank list obtained by our algorithm with the left top image as the most probable face for that person by our autonomous learning algorithm separately. We can see that among those top ranked faces, our algorithm can find the correct face for the person we are interested in, while Google may not.

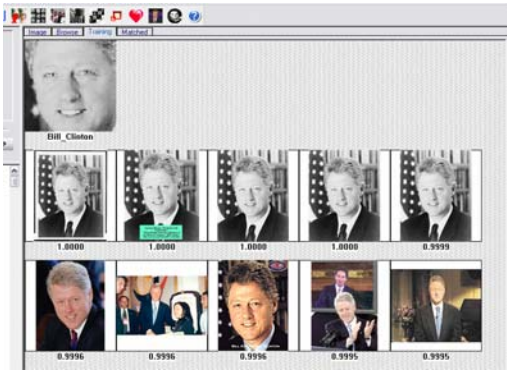
To compare the precision values, we manually annotated the ground truth, which is a huge work and thus limits the comparisons we can get. The images with profile faces and very small faces are all considered in the ground truth. Figure 6 and Table 1 show the precision and recall comparisons. We can see that even though we only extract the relatively big and frontal faces, which is not effective to those data with profile and very small faces, our algorithm still gets correct face models for those persons and improves the accuracy. For the case of "Bill Clinton", "Newt Gingrich", and "Hillary Clinton", we can get around 10% improvements on Average Precision [11] over the Google Image Search. For the case of "Madeleine Albright", where Google Search does a very good job and many profile and small faces occur, our average precision is still better.



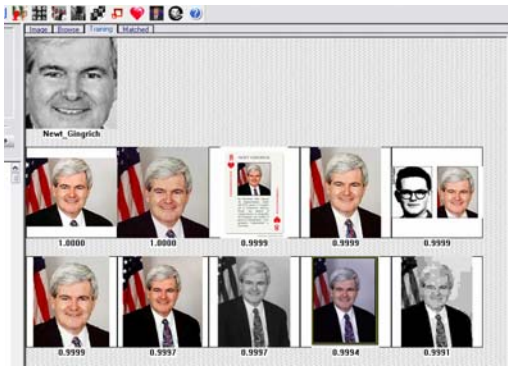
(a) "Newt Gingrich"

(b) "Hillary Clinton"

Figure 4. Illustration of Google Image Search Results

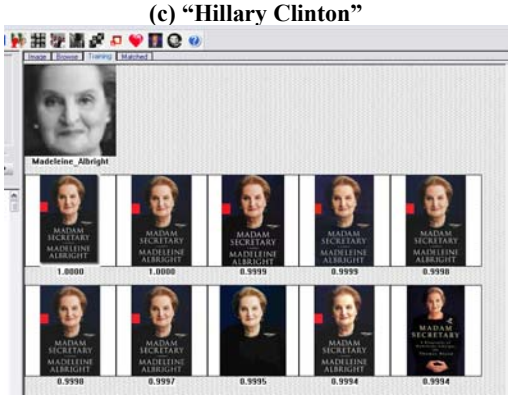


(a) "Bill Clinton"



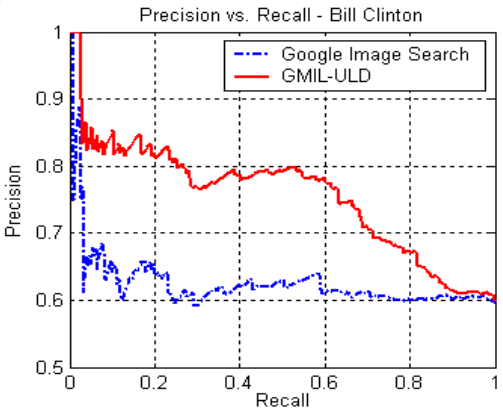
(b) “Newt Gingrich”

(c) “Hillary Clinton”

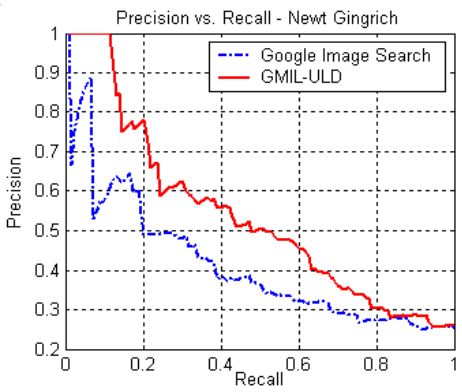


(d) “Madeleine Albright”

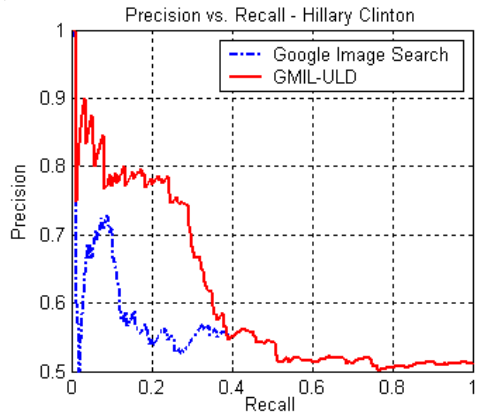
Figure 5. Illustration of the results by our algorithm



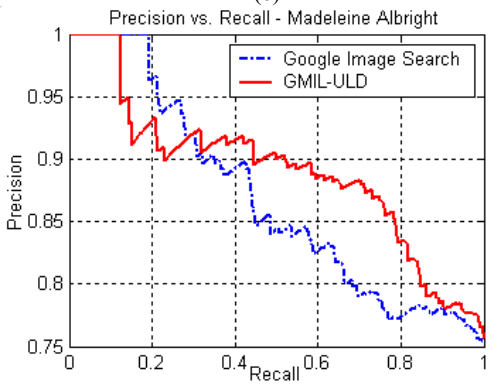
(a)



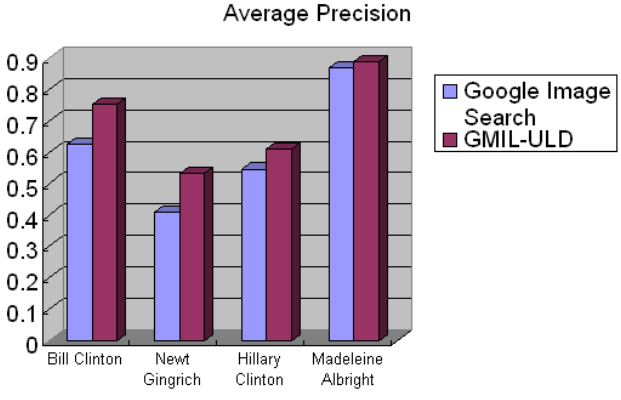
(b)



(c)



(d)



(e)

Figure 6. Performance comparison

Table 1: Comparison of Average Precision

Average Precision	Bill Clinton	Newt Gingrich	Hillary Clinton	Madeleine Albright
Google Image Search	0.6250	0.4100	0.5467	0.8683
GMIL-ULD	0.7546	0.5339	0.6107	0.8899

6. CONCLUSIONS

We have presented a cross-modality autonomous learning algorithm to build models for visual concepts based on image crawling from the results provided by search engines. Generalized MIL is proposed by introducing "Quasi-Positive Bags", and "Uncertain Labeling Density" is proposed to handle the Quasi-Positive Bags in order to find the most probable example for the concept we are interested in. Bag K-Means and Fuzzy Bag K-Means algorithms are proposed to find the maximum of DD and ULD respectively in an efficient way instead of the time-consuming gradient descent algorithm. The convergence of the algorithm is proved. Experiments are performed for learning the models for four persons. Comparing to Google Image Search results, our algorithm improves the accuracy and is able to build a correct model for a person. Ongoing works include applying this algorithm to learn more general concepts, e,g., outdoor and sports, as well as using these learned models for concept detection and search tasks in generic image/video databases, e.g., NIST TRECVID corpuses.

7. ACKNOWLEDGEMENT

We would like to thank Dr. Belle L. Tseng for her assistance on calculating average precision values in the experiments.

8. REFERENCES

[1]. X. Song and C.-Y. Lin and M.-T. Sun, Cross-modality automatic face model training from large video databases , The First IEEE CVPR Workshop on Face Processing in Video (FPIV'04), Washington DC, June 28, 2004

[2]. O. Maron, "Learning from ambiguity," PhD dissertation, Department of Electrical Engineering and Computer Science, MIT, Jun. 1998.

[3]. O. Maron, T. Lozano-Perez, "A Framework for Multiple Instance Learning," Proc. of Neural Information Processing Systems 10, 1998.

[4]. O. Maron, and A. L. Ratan, "Multiple-Instance Learning for Natural Scene Classification," Proc. of ICML 1998, 341-349.

[5]. T. G. Dietterich, R. H. Lathrop and T. Lozano-P_erez, "Solving the multiple instance problem with axis-parallel rectangles," Artificial Intelligence Journal, 89, 1997, 31-71.

[6]. R. A. Amar, D. R. Dooly, S. A. Goldman, and Q. Zhang, "Multiple-instance learning of real-valued data," Proc. of ICML, Williamstown, MA, 2001, 3-10.

[7]. J. Wang, and J. D. Zucker, "Solving Multiple-Instance Problem: A Lazy Learning Approach", Proc. of ICML, Stanford, CA, 2000, 1119-1125.

[8]. S. Andrews, T. Hofmann, and I. Tsochantaridis, "Multiple instance learning with generalized support vector machines, Proc. of the eighteenth national conference on Artificial Intelligence, Edmonton, Alberta, Canada, July 2002, 943 - 944

[9]. A. Schneider, "Weighted possibilistic clustering algorithms", Proc. of the 9th IEEE International Conference on Fuzzy Systems. Texas, 2000, 1, 176-180.

[10]. R.N. Dave, and R. Krishnapuram, "Robust clustering methods: a unified view", IEEE Transactions on Fuzzy Systems, May 1997, 5, 2, 270-293

[11]. A. Amir, M. Berg, S.-F. Chang, G. Iyengar, C.-Y. Lin, A. Natsev, C. Neti, H. Nock, M. Naphade, W. Hsu, J. R. Smith, B. Tseng, Y. Wu, D. Zhang, "IBM Research TRECVID-2003 Video Retrieval System," Proc. of TRECVID 2003 Workshop.

[12]. P. Viola and M. J. Jones, "Robust real-time object detection," Intl. J. Computer Vision, 2002.

[13]. R. Lienhart, A. Kuranov and V. Pisarevsky, "Empirical Analysis of Detection Cascades of Boosted Classifiers for Rapid Object Detection," DAGM-Symposium, 2003, 297-304.

[14]. M.J. Jones and J.M. Rehg, "Statistical color models with application to skin detection," Proc. of CVPR, 1999, 274-280.

[15]. Q. Zhang, and S. A. Goldman, "EM-DD: an improved multi-instance learning technique", Proc. of Advances in Neural Information Processing Systems, Cambridge, MA, MIT Press, 2002, 1073-1080.

[16]. C.-Y. Lin, B. L. Tseng and J. R. Smith, "Video Collaborative Annotation Forum: Establishing Ground-Truth Labels on Large Multimedia Datasets," *Proc. of NIST Text Retrieval Conf. (TREC)*, Gaithersburg, MD, November 2003.

[17]. C.-Y. Lin, B. L. Tseng, M. Naphade, A. Natsev and J. R. Smith, "VideoAL: A Novel End-to-End MPEG-7 Automatic Labeling System," *IEEE Intl. Conf. on Image Processing*, Barcelona, September 2003.

[18]. Y. Rui, T. 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, 8/5 (1998), 644-655.

[19]. P. Aigrain, H. Zhang, and D. Petkovic, "Content-Based Representation and Retrieval of Visual Media: A State-of-theArt Review," Multimedia Tools and Applications, Vol. 3, 179-202, November 1996.

[20]. T. Minka, "An Image Database Browser that Learns from User Interaction," MIT Technical Report TR#365, MIT, 1996.

[21]. S. Goldman, and Y. Zhou, "Enhancing Supervised Learning with Unlabeled Data," Proc of ICML, 2000.

[22]. Y. Wu, Q. Tian, T. S. Huang, "Discriminant-EM Algorithm with Application to Image Retrieval," Proc. of CVPR, Vol. I, pp. 222-227, Hilton Head Island, SC, June, 2000

[23]. G. Salton, and C. Buckle, "Improving retrieval performance by relevance feedback," Journal of the American Society for Information Science Vol. 41, 288-297, 1990

[24]. D. Lewis and W. A. Gale, "A sequential algorithm for training text classifiers," Proc. of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval, 1994.