Learning Representative Objects from Images Using Quadratic Optimization

Xiaonan Lu
Department of Computer Science
and Engineering
The Pennsylvania State University
Email: xlu@cse.psu.edu

Jia Li
Department of Statistics
The Pennsylvania State University
Email: jiali@stat.psu.edu

James Z. Wang
School of Information Sciences
and Technology
The Pennsylvania State University
Email: jwang@ist.psu.edu

***Abstract*— With the development of Content-Based Image Retrieval (CBIR) and ever increasing computing power, there is a notable growing interest in automatic learning from images. In this paper, we introduce a quadratic optimization based learning technique to enable computers to learn visual characteristics of a semantic concept from unlabeled images. In our work, images are represented by regions extracted from segmentation. Given a group of images conveying a semantic concept, we attempt to detect the region corresponding to the concept in every image using quadratic optimization. To characterize the visual properties of the concept, the mean of the feature vectors each describing the concept-associated region of an image is calculated and referred to as the representative feature vector. We apply the proposed learning technique to image classification and object recognition applications and provide experimental results.**

I. Introduction

Since the volume of image databases has been continuously increasing, it is imperative to advance automated image learning so that images can be effectively managed at the semantic level. Image semantic learning is a highly challenging problem in the crossroads of computer vision and machine learning.

Given a collection of pictures about a semantic concept, it is usually not difficult for human beings to find objects of interest and learn from images. For example, given five randomly selected images from a group of COREL images, shown in Figure 1, we are able to mark out "tiger" in every image and learn different kinds of visual characteristics of "tiger" from those pictures. It is known that the capability of learning and reasoning enables human beings to do so. Our aim is to equip computers with similar capability by developing automatic learning methodologies.

In this paper, we propose a quadratic optimization based learning technique to extract visual characteristics of semantic concepts from unlabeled images. For a group of images corresponding to a semantic concept, the proposed algorithm attempts to detect the regions in the images corresponding to the semantic concept and learn visual characteristics of the concept. For example, given the group of images in Figure 1, the goal of our algorithm is to automatically find in each image the area depicting the common object, i.e., the tiger, and extract the representative feature vector of tiger.

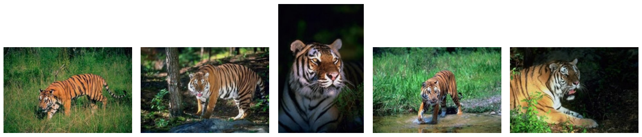

Fig. 1. Sample images about "tiger".

Potential applications of the proposed learning algorithm include object recognition and automatic image annotation. If we associate a textual description to a group of images to be learned, our algorithm will automatically link the textual description to the visual characteristics of the extracted object, useful for recognizing the object in new images. Since the proposed technique is able to link the semantic concept to visual characteristics of an object, it can also be applied in region-based image annotation.

A. Related Work

Content Based Image Retrieval (CBIR) has been an active research field since early 1990s. Many CBIR systems have been developed. IBM QBIC [6], [7], Virage [10], MIT Photobook [17], Columbia VisualSEEK and WebSEEK [23], [24], UCSB Netra [14], UIUC MARS [15], Berkeley Blobworld [3], UVA PicToSeek [9] and Stanford SIMPLIcity [27] are some of the systems. Rui et al. [19] published a survey on technical achievements of image retrieval in 1999. A recent article [22] by Smeulders provides review on the theory, techniques and applications of CBIR.

Aiming at improving performance of CBIR, automatic semantics categorization has been investigated. The Stanford SIMPLIcity [27] system classifies images into textured-nontextured, graph-photograph classes based on statistical methods. Chen et al. [4] extended the Multiple Instance Learning (MIL) algorithm and applied it to region-based image categorization. Another approach to improve CBIR is relevance feedback [21], [18], [5], an interactive learning technique. Optimization techniques have been exploited in relevance feedback [20].

Learning techniques have been applied to image retrieval for achieving higher performance. A review on statistical pattern recognition is referred to [12]. In [16], an unsupervised technique using eigenspace decomposition is developed for automatic object recognition. A retrieval architecture [26] is proposed to adjust the interrelationship between feature selection, feature representation, and similarity metric. A learning method that extracts heterogeneous models of object classes for visual recognition is proposed in [28]. A search principle for optimal feature subset selection using the Branch and Bound method is introduced in [25].

There are systems which associate textual information to images based on the features of regions [2], [1] or of the entire image [13]. The work [2], [1] at UC Berkeley presents a region-based approach to automatically annotate images. In their approach, an image is represented as a set of regions. Several correspondence models about image regions and words are trained to find the joint distribution of regions and words. Based on the trained correspondence model, images are matched to words, or vice versa. The ALIP system developed by Li et al. [13] automatically assigns textual description to images using a statistical modeling approach. For every concept, feature vectors of training images are profiled by two-dimensional multi-resolution hidden Markov model. Given a test image, the likelihood between its feature vectors and each stored trained model is computed. A small set of statistically important index terms about the image are given based on the calculated likelihoods.

B. Our Approach

In our work, an image is represented by a set of regions. Every image is segmented into regions, each roughly homogeneous in color and texture. For each region, a feature vector describing the average color, texture, shape and area percentage of its coverage area is calculated. Finally, an image is represented by a set of feature vectors corresponding to regions in the image. Heuristically, regions in an image correspond to objects in real world under high-quality segmentation.

We attempt to design a learning algorithm for computers to detect the representative object and learn visual characteristics of the object automatically from a group of images. By looking at the five images about "tiger" in Figure 1, we make the following observations about those images:

- There is a region with similar color (brown), texture and shape in every image;
- Each of the visually similar regions occupies a considerably large area.

It is reasonable to assume that for a particular concept, every training image contains an object corresponding to the concept. Also, the object should be an important part in each training image, that is, the corresponding region should occupy a considerably large area in every image. We also assume that images of the same object share similar visual characteristics. Based on these assumptions, we convert the problem of detecting the representative object for a semantic concept to a new problem of *finding regions, one from each image, that not only are mutually similar in terms of color, texture and shape but also individually occupy a considerably large area in their corresponding images.*

The proposed learning algorithm proceeds in two steps. In the first step, the algorithm aims at detecting the region corresponding to the representative object in every training image using quadratic optimization. In the second step, a representative feature vector is computed using the features of the identified regions. This representative feature vector reflects the visual characteristics of the common object in the training images.

For a group of images corresponding to a concept, our algorithm not only learns the visual characteristics of the concept but also finds the corresponding region in each image, which is desirable for many applications. In addition, since our algorithm learns each concept independently, training new concepts or updating training images of a trained concept does not incur computation outside the domain of the corresponding concept, entailing good scalability.

C. Outline of the Paper

The remainder of the paper is organized as follows: Section 2 describes the region based image representation method and the similarity metric. In Section 3, the quadratic based learning algorithm is presented. In Section 4, applications of our learning algorithm and experimental results are presented. Finally, we conclude and suggest future research in Section 5.

II. Image Representation

In our work, every image is represented by a set of regions obtained from image segmentation. In this section, the image segmentation and feature extraction methods are presented. The definition of similarity metric is provided subsequently.

A. Image Segmentation and Feature Extraction

For each image, the system first partitions the image into blocks of 4×4 pixels. The reason for this partition is to reduce the computation cost for image segmentation. The block size is chosen to trade off image details and computation time. For each block, a feature vector describing its average color and texture is extracted. Then, the k-means algorithm [11] is used to cluster the feature vectors into several classes with every class corresponding to one "region" in the segmented image.

In the feature vector extracted from each block, there are 6 features. Three of them are the average color components of the 4×4 block. The other three features characterize texture of the block. For color representation, we use the well-known LUV color space, where L encodes luminance and U and V encode color information(chrominance). The LUV color space is chosen because of its good perception correlation properties. To obtain the other three features for texture, we apply a Daubechies-4 wavelet transform to the L component of the image. After a one-level wavelet transform, a 4×4 block is decomposed into four frequency bands: the LL, LH, HL, and HL bands. Each band contains 2×2 coefficients. Without loss of generality, we suppose the coefficients in the HL band are $\{c_{k,l}, c_{k,l+1}, c_{k+1,l}, c_{k+1,l+1}\}$. One feature is

$$\left(\frac{1}{4} \sum_{i=0}^{1} \sum_{j=0}^{1} c_{k+i,l+j}^{2} \right)^{\frac{1}{2}} \tag{1}$$

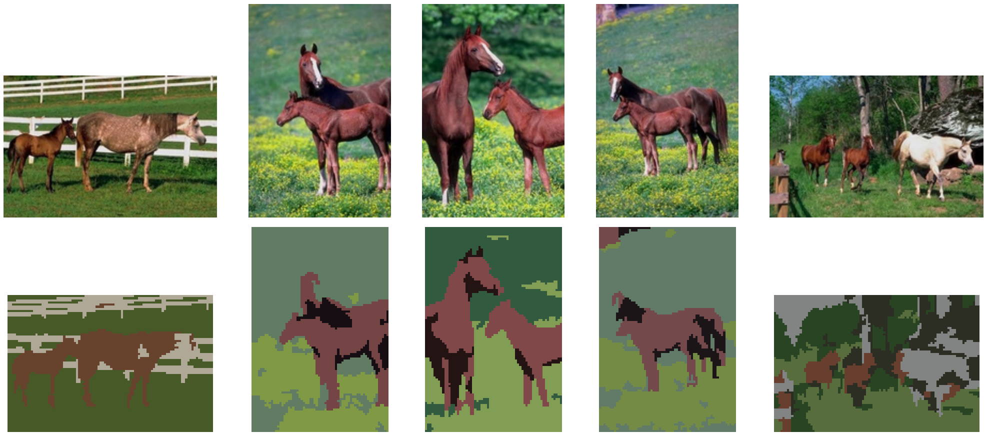

Fig. 2. Segmentation result for images about "horse". The first line: original images. The second line: images after segmentation.

The other two features are computed in the same way from coefficients of LH and HH bands. The HL band reflects activities in the horizontal direction. A local texture of vertical strips has high energy in HL band and low energy in LH band.

Suppose an image is represented by a set of feature vectors $\{x_i : i = 1, \cdots, L\}$, the goal of the k-means algorithm is to partition the set of feature vectors into k groups with means $\hat{x_1}, \hat{x_2}, \cdots, \hat{x_k}$ such that

$$D(k) = \sum_{i=1}^{L} \min_{1 \le j \le k} (x_i - \hat{x_j})^2 \tag{2}$$

is minimized. The algorithm does not specify the number of clusters, k, to choose. We adaptively choose the number of clusters by gradually increasing k and stop when a stopping criterion is met. We do not give detailed description of the stopping criteria here due to space limit. Readers are suggested to find it in [27].

After segmentation, the mean of the feature vectors in each cluster is used to represent color and texture attributes of the corresponding region. In addition, three extra features describing shape properties are calculated using normalized inertia [8]. For a region r_j in the image, its normalized inertia of order γ is given as:

$$I(r_j, \gamma) = \frac{\sum_{x : x \in r_j} \| x - \hat{x_j} \|^{\gamma}}{v_j^{1+\gamma/2}} \tag{3}$$

where $\hat{x_j}$ is the centroid of r_j, v_j is the number of pixels in region r_j. The normalized inertia is invariant to scaling and rotation. The minimum normalized inertia is achieved by spheres. Denote the γth order normalized inertia of spheres as L_{γ}. We define three shape features as:

$$s_{j\gamma} = I(r_j, \gamma) / L_{\gamma} \quad \gamma = 1, 2, 3 \tag{4}$$

We would also like to describe area percentage of a region r_j, which is the percentage of the image covered by region r_j

$$q_j = \frac{v_j}{V} \tag{5}$$

where v_j is the number of pixels in the region r_j and V is the number of pixels in the image. Finally, a region is represented by a ten-dimensional feature vector, with three color elements, three texture elements, three shape elements and one area percentage element. An image is represented by a set of feature vectors with each corresponding to one region.

To demonstrate the image segmentation result, images and their segmented regions are shown in Figure 2. Every region obtained from image segmentation is shown by its average color component.

B. Similarity Metric

In our system, distance between two images is defined as the weighted sum of distances between regions. For two images m, m', if we let k and k' be the number of regions in m and m' respectively, distance between these two images is defined as:

$$d(m, m') = \sum_{i=1}^{k} \sum_{j=1}^{k'} w_{ij} d(r_i, r'_j) \tag{6}$$

where w_{ij} is the weight for a region pair r_i, r'_j, $d(r_i, r'_j)$ is the distance between two regions.

For two regions r and r', let their feature vectors be f and f'. We denote the distance between the two regions by $d(f, f')$ or $d(r, r')$ for the sake of stressing between-region distance. This distance is defined by

$$d(r, r') = g(d_s(r, r')) \cdot d_t(r, r') \tag{7}$$

where $d_t(r, r')$ is the color and texture distance defined as:

$$d_t(r, r') = \sum_{i=1}^{6} w_i (f_i - f'_i)^2 \tag{8}$$

and $d_s(r, r')$ is the shape distance computed by

$$d_s(r, r') = \sum_{i=7}^{9} w_i (f_i - f'_i)^2 \tag{9}$$

The function $g(d_s(r, r'))$ is a nonlinear converting function to ensure a proper influence of the shape distance on the total distance. Readers are suggested to find the detailed description of these functions in [27].

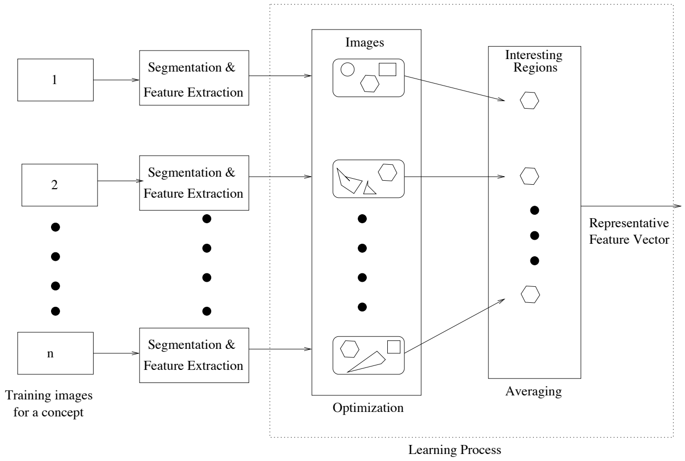
Images
Interesting Regions
1
Segmentation & Feature Extraction
2
Segmentation & Feature Extraction
Representative Feature Vector
n
Segmentation & Feature Extraction
Training images for a concept
Averaging
Optimization
Learning Process

Fig. 3. Detecting and learning representative object.

III. Quadratic Optimization Based Learning

In this section, we present in detail the quadratic optimization based learning technique. The learning process, rational for the learning method and formulation of the quadratic optimization problem are presented.

A. Learning process

Given a collection of images about a concept, the process of detecting representative object and learning visual characteristics of the object is illustrated in Figure 3. During the image segmentation and feature extraction step, every image is represented by a set of regions. Each region is described by a ten-dimensional feature vector, containing color, texture, shape and area percentage information of the region. Regions in segmented images correspond to data points in the ten-dimensional feature space, and the entire set of data points is the input to the learning process. During the learning process, the representative object in each image is detected by quadratic optimization. In the optimization framework, each region is assigned a weight value which indicates significance of the region within its embedding image. If a region has higher weight value, it has more influence on determining the semantics of its embedding image. After the optimization of weights, the region with the maximum weight value in every image is selected. As the collection of images conveys a specific semantic concept, the selected regions represent visual expressions of the representative object corresponding to the semantic concept. In the feature space, the centroid of the selected regions is learned as the representative feature vector for the semantic concept. For a group of images corresponding to a semantic concept, representative object is detected and representative feature vector is learned through three steps shown in Figure 3.

B. Optimization Method

Even though the area percentage is commonly used to represent significance of an image region, it only reflects the relationship between a region and the embedding image. We need a method that assigns a weight to a region based on the whole group of training images corresponding to a semantic concept. In an attempt to detect representative object, we need to analyze the whole group of training images and find regions that are recommended globally by the images. Assuming that images of a same object share similar visual characteristics, we seek for similar regions across the entire training image set.

An integrated region matching similarity metric is used to measure similarity between a pair of images. Similarity between two images is a weighted sum of distances between regions in two images. Regions with higher weight parameter values have more influence on determining similarity between two images. While a similarity metric measures distance between two images, the sum of pairwise distances between images reflects similarity of a group fo images. It is obvious that if higher weight value is given to the representative object in every image, the sum of pairwise distances between training images tens to be small. Because higher weight values are given to visually similar regions, which have lower between-region distance value, the between-image distances become small. In the reverse way, visually similar regions across the entire training image set would get higher weight values by optimizing the sum of pairwise distances between images.

C. Formulation of the quadratic optimization problem

We formulate the task of assigning weight values to regions as a quadratic optimization problem. The objective is to minimize the sum of pairwise distances between images. To facilitate presentation, we first introduce notations used in this section:

- n represents the number of images in a group. Id of an image is within $\{1, 2, \cdots, n\}$.
- m_i represents the number of regions in image i.
- $v_k^{(i)}$ represents feature vector of region k in image i.
- $q_k^{(i)}$ represents area percentage of region k in image i.
- $p_k^{(i)}$ represents significance weight of region k in training image i.

The distance between two images is calculated according to Equation (6). For any pair of regions, for example region k in image i and region l in image j, the weight in the image similarity metric assigned to this pair is the multiple of their area percentages. The distance between image i and image j is calculated as:

$$\sum_{k=1}^{m_i} \sum_{l=1}^{m_j} p_k^{(i)} p_l^{(j)} d(v_k^{(i)}, v_l^{(j)})$$

where the weight w_{kl} for region k in image i and region l in image j is : $w_{kl} = p_k^{(i)} p_l^{(j)}$. The distance between two regions is represented by $d(v_k^{(i)}, v_l^{(j)})$.

The quadratic optimization problem for detecting representative object is formulated as the following:

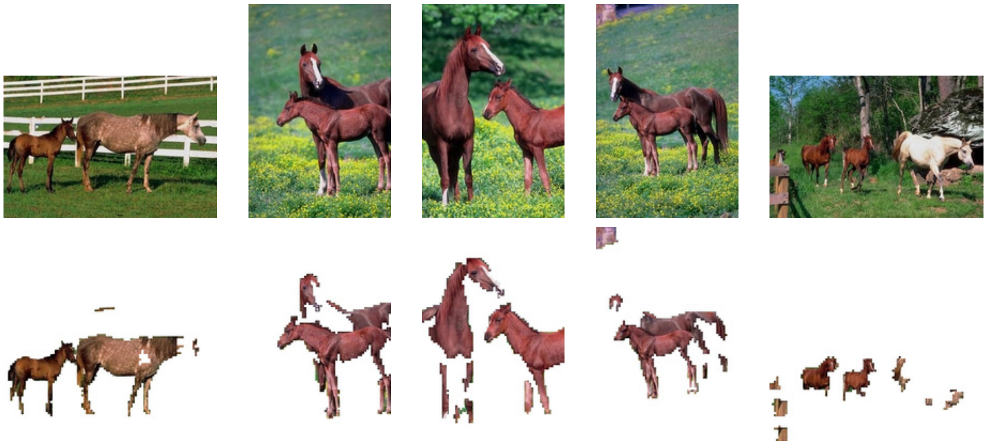

Fig. 4. Representative object of "horse" images. First line: original images. Second line: images highlighting detected representative object.

Minimize D+P

$$D = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\sum_{k=1}^{m_i}\sum_{l=1}^{m_j} p_k^{(i)} p_l^{(j)} d(v_k^{(i)}, v_l^{(j)})\right)$$

$$P = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} \lambda \left[\sum_{k=1}^{m_i}\left(p_k^{(i)} - q_k^{(i)}\right)^2 + \sum_{l=1}^{m_j}\left(p_l^{(j)} - q_l^{(j)}\right)^2\right]$$

subject to:

$$\sum_{k=1}^{m_i} p_k^{(i)} = 1 \qquad i = 1, 2, \cdots, n$$

$$p_k^{(i)} \geq 0 \qquad i = 1, 2, \cdots, n \quad k = 1, 2, \cdots, m_i \; .$$

Variables of this quadratic optimization problem are weight parameters of regions: $p_k^{(i)}$, $i = 1, 2, ..., n$, $k = 1, 2, ..., m_i$. Constraints of the problem state that weight parameter should be nonnegative and the sum of weight parameters of regions in every image equals one.

Objective function of the quadratic optimization problem is the sum of two parts: D and P. D is the sum of pairwise distances between images. P represents the total penalty, which is defined as the sum of penalties for every pair of images. Penalty for every two images is defined as the multiple of a penalty parameter λ and the difference between weight values and area percentages. λ is an adjustable non-negative parameter which reflects the influence of area percentage on weight. If $\lambda = 0$, there is no penalty in the objective function. Area percentage of a region has no effect on determining the weight of the region. On the other hand, if λ is big enough, total penalty is predorminant in the objective function. Weight of a region is completely decided by the area percentage, which also means that the weight equals area percentage. The purpose of an added penalty in the objective function is to prevent the algorithm from finding a pretty small object which happens to be visually similar across the entire training images. Since we assume that the representative object should occupy a substantial area in every image, the small region which has the highest weight could not be the objective.

In real system, appropriate choice of λ depends on the image segmentation result. If every region tends to occupy a large area, which means that there is little probability that a very small object alone corresponds to a region, λ is set to zero or a small value. On the other hand, if the average number of regions in an image is large and there is a high chance that a very small object corresponds to a region, fine tuning λ is critical for the performance.

D. Learning Representative Feature Vector

Representative feature vector of a semantic concept is estimated from regions corresponding to the representative object. For a set of training images corresponding to a semantic concept, solution of the proposed quadratic optimization problem contains weight for every region. In every training image, the region with the highest weight value is selected as the region corresponding to the representative object. Having the set of selected regions, which represent images of the representative object, mean of feature vectors is regarded as the representative feature vector for the trained semantic concept.

IV. Experiments

We implement an experimental system using C programming language and MATLAB optimization toolbox. For a group of training images corresponding to a semantic concept, the entire set of regions is obtained from image segmentation. MATLAB optimization toolbox is used to do quadratic optimization on weights of regions. In every image, the region with the highest weight is selected as the representative object. The centroid of the selected regions is learned as the representative feature vector for the concept. We applied the proposed learning technique on two datasets. One is a subset of "horse" in COREL database. The other is a dataset of images containing traffic signs.

A. Detecting Representative object

Since the average number of regions in an image is small using the proposed image segmentation technique (between 3 and 4 during our test), each region occupies a considerable area

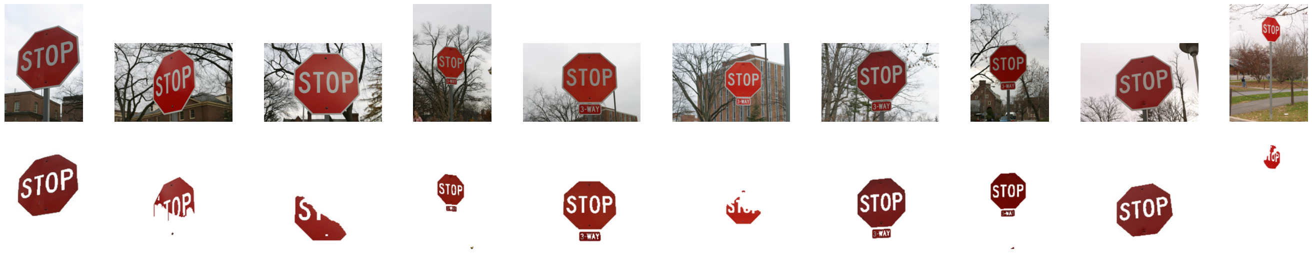

Fig. 5. Representative object of stop sign training images. First line: original images. Second line: images highlighting detected representative object.

percentage. We set the penalty parameter λ as zero because there is little probability that a very small object will be detected. To visualize the representative object learned by our technique, we implemented a tool which highlights the region detected as representative object in every training image. In the figures which show detected representative object, the region with highest weight in every training image is shown in its original color, while other regions are shown in pure white. We selected five samples from "horse" images in COREL database as training set and applied the proposed learning method. In every image, the region corresponding to the detected representative object is shown in Figure 4. As an application of object recognition, we use a dataset of 216 images targeting traffic signs. 50 images of the dataset contains stop signs in different background and angel. We use 10 stop sign images as the training set for detecting representative object, stop sign, and learning representative feature vector. In every training image, the detected representative object is shown in Figure 5.

B. Learning Visual Characteristics of Stop Sign

Having detected representative object in every stop sign training image, the mean of the feature vectors associated with stop sign is calculated and referred to as the representative feature vector. To estimate the distribution of detected representative objects in training images, empirical probability function (pdf) and cumulative distribution function (cdf) of distances between the representative feature vector and each feature vector describing the stop sign region are shown in Figure 6.

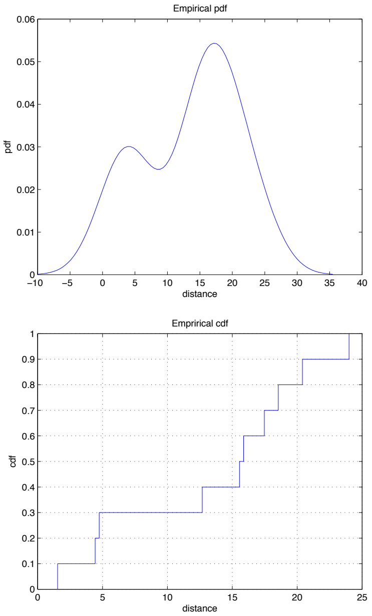

Fig. 6. Empirical probability density function and cumulative density function of the distances from training images to the representative feature vector.

C. Stop Sign Recognition

As an application of the proposed learning technique, stop sign recognition is tested. The training set contains 10 stop sign images, and the representative feature vector is obtained using the quadratic optimization. The test set is the entire self-prepared dataset, which contains 216 images belonging to various categories: indoor, outdoor, human, stop sign, road, traffic, scene, classroom etc.

Stop sign images are recognized based on distance between each test image and the learned representative feature vector of stop sign. The distance between a test image and the representative feature vector is defined as the minimum distance between any region of the test image and the representative feature vector. To estimate the distribution of distances between test images and the representative feature vector, empirical probability function (pdf) and cumulative distribution function (cdf) of calculated distances are shown in Figure 7. A specific threshold distance value is also needed for stop sign recognition. Every time we deal with a test image, we compare its distance to the representative feature vector with the threshold distance value. If the calculated distance is below the threshold distance value, the image is labeled as containing a stop sign. Otherwise, the image is regarded as not containing any stop sign. Precision and recall of the stop sign recognition under different threshold distance values are shown in Figure 8.

D. Effect of Image Segmentation

Since the proposed learning technique uses region-based approach, it is sensitive to correctness of image segmentation. In order to investigate the effect of image segmentation algorithm on the performance of object recognition, we tested stop sign recognition under a different setting. By manual inspection, we found that even though there are 50 images in the test set containing stop sign, there are only 34 of them having regions roughly corresponding to the stop sign object after

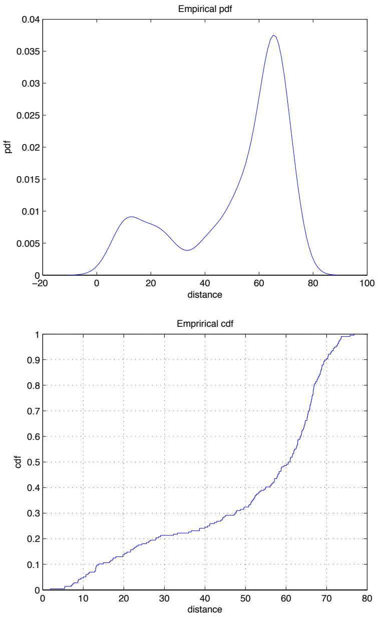

Fig. 7. Empirical probability density function and cumulative density function of the distances from test images to the representative feature vector.

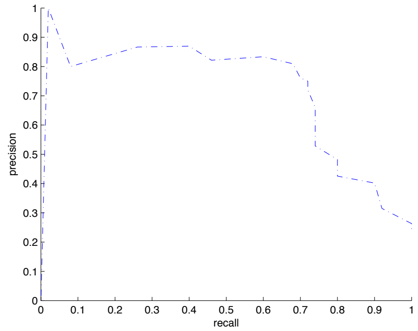

Fig. 8. Precision-recall curve for stop sign recognition.

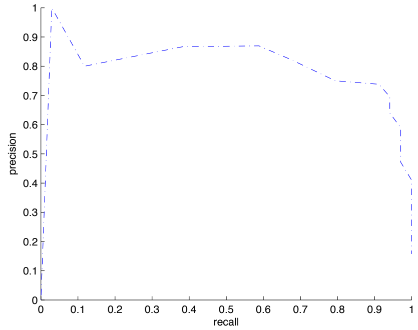

Fig. 9. Precision-recall curve for stop sign recognition under the simulation of high quality image segmentation.

image segmentation. In an attempt to simulate the condition of high quality image segmentation, we regard these 34 images as the only ones that contain stop signs when we calculate precision and recall. Performance under this new setting is shown in Figure 9. Comparing Figure 8 and Figure 9, we see that performance of object recognition using the proposed learning technique can be considerably improved by better segmentation.

V. Conclusions and Future Work

In this paper, we have presented a quadratic optimization based learning technique to detect representative objects from images and extract visual characteristics of the object automatically. For a group of images from a semantic concept, the region associated with the concept in every image is detected by quadratic optimization. Visual characteristics of the concept are learned from the detected regions. Preliminary experiments on detecting representative object and object recognition are presented. Several issues need to be investigated in greater depth in future work. First, the image segmentation method can be improved. Second, in the current work, we assumed that the object of interest corresponds to one region in each image, which may not be true especially due to the difficulty of segmentation. A method to extract multiple significant regions needs to be developed. Third, the distance measure between regions can be refined.

VI. Acknowledgment

The material about the SIMPLIcity system was based upon work supported by the National Science Foundation under Grant No. IIS-9817511 and in part by Stanford University. The work presented in this paper is supported by the National Science Foundation under Grant Nos. IIS-0219272 and IIS-0333036, The Pennsylvania State University, the PNC Foundation, and SUN Microsystems.

References

[1] K. Barnard, P. Duygulu, N. Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] K. Barnard and D. Forsyth. Learning the semantics of words and pictures. In *Int. Conf. Computer Vision*, pages 408–415. IEEE, 2001.

[3] C. Carson, S. Belongie, H. Greenspan, and J. Malik. Region-based image querying. *Workshop on Content-Based Access of Image and Video Libraries*, 1997.

[4] Y. Chen and J. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, Aug 2004.

[5] I. J. Cox, M. L. Miller, S. M. Omohundro, and P. N. Yianilos. Target testing and the pichunter bayesian multimedia retrieval system. In *Int. Forum on Research and Technology Advances in Digital Libraries*, page 66. IEEE, 1996.

[6] C. Faloutsos, R. Barber, M. Flickner, J. Hafner, W. Niblack, D. Petkovic, and W. Equitz. Efficient and effective querying by image content. *Journal of Intelligent Information Systems*, 3(3/4):231–262, 1994.

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

[8] A. Gersho. Asymptotically optimum block quantization. *IEEE Trans. Inform. Theory*, 25(4):373–380, Jul 1979.

[9] T. Gevers and A. Smeulders. Pictoseek: Combining color and shape invariant features for image retrieval. *IEEE Trans. Image Processing*, 9(1):102–119, 2000.

[10] A. Gupta and R. Jain. Visual information retrieval. *Communications of the ACM*, 40(5):70–79, May 1997.

[11] J. A. Hartigan and M. A. Wong. Algorithm as136: a k-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[12] A. K. Jain, R. R. W. Duin, and J. Mao. Statistical pattern recognition: A review. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(91):4–37, 2000.

[13] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9):1075–1088, 2003.

[14] W.-Y. Ma and B. S. Manjunath. Netra: A toolbox for navigating large image databases. *Multimedia Systems*, 7(3):184–198, 1997.

[15] S. Mehrotra, Y. Rui, M. Ortega, and T. S. Huang. Supporting content-based queries over images in MARS. *Int. Conf. Multimedia Computing and Systems*, pages 632–633, 1997.

[16] B. Moghaddam and A. Pentland. Probabilistic visual learning for object representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 19(7):696–710, 1997.

[17] A. Pentland, R. Picard, and S. Sclaroff. Photobook: Content-based manipulation of image databases. *International Journal of Computer Vision*, 18(3):233–254, 1996.

[18] R. W. Picard. Digital libraries: Meeting place for high-level and low-level vision. In *Recent Developments in Computer Vision*, pages 3–12. Springer, 1995.

[19] Y. Rui, T. Huang, and S. Chang. Image retrieval: current techniques, promising directions and open issues. *Journal of Visual Communication and Image Representation*, 10(4):39–62, April 1999.

[20] Y. Rui and T. S. Huang. Optimizing learning in image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 236–243. IEEE, 2000.

[21] Y. Rui, T. S. Huang, and S. Mehrotra. Content-based image retrieval with relevance feedback in mars. In *IEEE Conf. on Image Processing*. IEEE, 1997.

[22] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[23] J. R. Smith and S.-F. Chang. Visualseek: A fuly automated content-based image query system. In *ACM Multimedia*, pages 87–98. ACM, July 1996.

[24] J. R. Smith and S.-F. Chang. Visually searching the web for content. *IEEE Multimedia*, 4(3):12–20, July 1997.

[25] P. Somol, P. Pudil, and J. Kittler. Fast branch and bound algorithm for optimal feature selection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(7):900–912, July 2004.

[26] N. Vasconcelos and A. Lippman. A probabilistic architecture for content-based image retrieval. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 216–221. IEEE, 2000.

[27] J. Z. Wang, J. Li, and G. Wiederhold. Simplicity: Semantics-sensitive integrated matching for picture libraries. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(9):947–963, 2001.

[28] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 101–108. IEEE, 2000.