ASPECT MODELING OF PARSED REPRESENTATION FOR IMAGE RETRIEVAL

Soo Hyun Bae and Biing-Hwang Juang

Center for Signal and Image Processing
Georgia Institute of Technology, Atlanta, GA 30332
{soohyun,juang}@ece.gatech.edu

ABSTRACT

A probabilistic framework based on a universal source coding for content-based image retrieval is proposed. By a multidimensional incremental parsing technique, which is an extension of the Lempel-Ziv incremental parsing algorithm, a given image is parsed into a number of variable-size rectangular blocks, called parsed representations. To achieve a semantically relevant pattern matching, we introduce a new similarity measure from the first- and second-order statistics of given image patches. Once the occurrence patterns of images in the corpus are analyzed, the term-document joint distribution is estimated by an aspect modeling technique under the assumption of latent aspects. To compare the performance of the proposed image retrieval framework based on the parsed representations, we implement a benchmark system based on the fixed-shape block representations trained by vector quantization. In addition to these two systems, we bring two content-based image retrieval systems into the performance evaluation. The experimental results on a database of 20,000 natural scene images demonstrate that the proposed image retrieval system significantly outperforms other existing and the benchmark systems.

Index Terms— Image Retrieval, Pattern Recognition, Incremental Parsing, Latent Semantic Analysis

1. INTRODUCTION

Since Lempel and Ziv proposed a dictionary-based source coding algorithm in [1], there have been a plethora of research on the multidimensional extensions of the algorithm. However, it was not clear that any of them can deal with a given multidimensional source without the aid of a scanning scheme. This previously motivated us to devised a multidimensional incremental parsing scheme for universal source coding [2]. The proposed scheme parses a given multidimensional source into a number of variable-size patches, we call this methodology a parsed representation, and achieves a significantly improved performance of image compression applications. Based on the source characterization property of the proposed parsing scheme, we also proposed a query-by-example content-based image retrieval (CBIR) framework [3] and implemented image retrieval systems, called IPSILON systems. The proposed CBIR systems analyze the semantic concepts of visual information by the latent semantic analysis (LSA). An extensive comparison of the IPSILON systems with other existing systems experimentally showed the superior performance of the proposed CBIR framework. However, there exist several fundamental limitations of the framework. First, although the fundamental assumption of LSA is that words and documents form a joint Gaussian distribution, it is known that Poisson or negative binomial distribution is more appropriate for term count. Second, since the approximate co-occurrence matrix is a Gaussian distribution, it may contain negative entries for occurrence counts, which is obviously an unsuitable approximation for term counts. Third, LSA is not readily extensible to a more flexible framework that can necessarily deal with heterogeneous sources.

To overcome the aforementioned drawbacks, in this paper, we propose a probabilistic image retrieval framework for query-by-example CBIR systems. In visual information analysis applications, e.g., image retrieval and annotation, a considerable number of techniques that take advantage of the aspect modeling technique by probabilistic latent semantic analysis (PLSA) [4] have been reported in literature. Most of them extract visual features from fixed-block partition or image segments. The proposed framework uses the dictionary entries of the incremental parsing as features of the given image. Once the co-occurrence matrix of a given image corpus is generated, we train the aspect model by PLSA and design image retrieval systems, called AMPARS (Aspect Modeling of PArsed RepreSentation). Another difference of AMPARS compared with the previous systems is the pattern matching scheme. IPSILON use a pixel-wise minimax distortion function for evaluating the perceptual similarity of patches. However, a desirable attribute of pattern matching in image retrieval systems is to minimize semantic discrepancy instead of perceptual distortion. To accomplish this, it is necessary to consider distortion metrics that go beyond pixel-by-pixel comparisons, e.g., by considering first- and second-order statistics of given patches. A promising early step in this direction is the development of a new class of structural similarity (SSIM) metrics by Wang *et al* [5]. SSIM metrics have been shown to be robust to various types of image perturbations that are not sensitive to the human eyes. However, the effective use of SSIM metrics in retrieval applications require significant adaptation, e.g., complete reliance on region statistics. In this paper, we adopt the structural texture similarity (STSIM) metric proposed in [6]. Since they were originally desined for comparison of gray-scale images, we propose a variation of STSIM for the measurement of color image patches.

To compare the effectiveness of the use of the dictionary entries by incremental parsing (IP) under PLSA framework, we implemented a benchmark retrieval system that uses a visual dictionary trained by vector quantization (VQ). Also, the performance of these systems is compared with that of two other systems: the IPSILON system and the SIMPLIcity proposed by Wang *et al.* [7]. The performance of all these four systems are evaluated with a database of 20,000 images of natural scenes. The results indicate that the proposed framework offers considerable performance improvements over the other three techniques.

This work was supported in part by Hewlett-Packard Laboratories. The authors are grateful to James Wang for providing the implementation of the SIMPLIcity.

2. PROBABILISTIC LATENT SEMANTIC ANALYSIS

Suppose we have a collection of N documents $D = \{d_1, \cdots, d_N\}$ and a lexicon with M words $W = \{w_1, \cdots, w_M\}$. Based on the vector space model, a document is represented as an M-dimensional vector. From the observation of a given corpus, we can compute the empirical distribution $p(w, d)$ that corresponds to the term-document co-occurrence in LSA. Basically, PLSA estimates the term-document joint distribution $P(w, d)$ that minimizes the Kullback-Leibler (KL) divergence with respect to the empirical distribution $p(w, d)$ subject to K latent aspects, as opposed to the L_2 norm minimization performed by LSA. In aspect modeling, each document is a mixture of latent aspects $z_k \in Z = \{z_1, \cdots, z_K\}$.

PLSA model has two independence assumptions. First, observation pairs (w_i, d_j) are generated independently. Second, the pairs of random variables (w_i, d_j) are conditionally independent given the hidden aspect z_k, i.e.,

$$P(w_i, d_j|z_k) = P(w_i|z_k)P(d_j|z_k). \tag{1}$$

The joint distribution of the observation pairs is the marginalization over the K latent aspects z_k as follows:

$$P(w_i, d_j) = P(d_j) \sum_{z_k \in \mathcal{Z}} P(w_i|z_k)P(z_k|d_j). \tag{2}$$

The estimation of the conditional probability distributions $P(w_i|z_k)$ and $P(z_k|d_j)$ can be resolved by applying the expectation-maximization (EM) technique, which maximizes the likelihood function of the observed data

$$\mathcal{L} = \prod_{j=1}^{N} \prod_{i=1}^{M} P(d_i) \sum_{k=1}^{K} P(z_k|d_j)P(w_i|z_k)^{p(w_i,d_j)}. \tag{3}$$

The output of the EM algorithm under PLSA is the two multinomial distributions $P(w|z)$ and $P(z|d)$, from which the joint distribution $P(w, d)$ is estimated.

3. STRUCTURAL SIMILARITY INDEX

One of the recently proposed class of image fidelity measures is the structural similarity (SSIM) [5], which is not based on explicit models of the human visual system (HVS) or measurements of noise sensitivities, but instead, account for higher-level functionalities of the HVS, and in particular, make use of the fact that it can extract structural information in the form of relative spatial covariance from the viewing field.

There are several SSIM implementations, both in image domain and the wavelet domain. The basic SSIM metric presented in [5] is a real number in the range [-1,1] and is computed based on the second-order statistics of the reference and the distortion images as follows:

$$\text{SSIM}(\mathbf{x}, \mathbf{y}) = \frac{(2\mu_x\mu_y + C_1)\,(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)\,(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{4}$$

where $\mathbf{x}$ and $\mathbf{y}$ are two nonnegative image signals, μ_x and μ_y are the mean intensities, σ_x^2 and σ_y^2 are the variances, σ_{xy} is the covariance of the $\mathbf{x}$ and $\mathbf{y}$, and C_1 and C_2 are small real constants relative to the μ_x and μ_y. The spatial domain SSIM has been shown to provide good quality prediction across a variety of artifacts, but is highly sensitive to spatial translation.

Although the above SSIM indexes primarily focus on comparing the structural information from the images, it has not been used in texture comparison applications since the original SSIM is too constrained to capture the perceptual similarity of two textures. Zhao *et al.* in [6] replaced the structure term with structural texture terms that are sensitive to local textual statistics. The first-order autocovariance in the horizontal direction is defined as

$$\rho_x(0, 1) = E\{(x_{i,j} - \mu_x)(x_{i,j+1} - \mu_x)\}/\sigma_x^2. \tag{5}$$

The autocovariance in the vertical direction is defined in a similar fashion. The texture term in the horizontal direction is

$$c_{0,1}(\mathbf{x}, \mathbf{y}) = 1 - 0.5(|\rho_x(0, 1) - \rho_y(0, 1)|)^p. \tag{6}$$

In [6], p is set to 1. These horizontal and vertical texture terms are combined with the luminance and the contrast terms in the original SSIM as follows:

$$\text{STSIM}(\mathbf{x}, \mathbf{y}) = l(\mathbf{x}, \mathbf{y})^{\frac{1}{4}} c(\mathbf{x}, \mathbf{y})^{\frac{1}{4}} c_{0,1}(\mathbf{x}, \mathbf{y})^{\frac{1}{4}} c_{1,0}(\mathbf{x}, \mathbf{y})^{\frac{1}{4}}, \tag{7}$$

where $l(\mathbf{x}, \mathbf{y}) = (2\mu_x\mu_y + C_1)/(\mu_x^2 + \mu_y^2 + C_1)$ and $c(\mathbf{x}, \mathbf{y}) = (2\sigma_x\sigma_y + C_2)/(\sigma_x^2 + \sigma_y^2 + C_2)$.

However, there has been no structural similarity measure that takes into account color images or image patches in image retrieval applications. Thus, we here propose a color structural texture similarity measure (CTSIM) for a matching criterion of the universal source coding.

Let $\mathbf{x}$ and $\mathbf{y}$ be two images or image patches to be compared represented in three color components, i.e., red, green, and blue (RGB). When each image patch is represented in YCbCr domain as $\mathbf{x} = \{\mathbf{x}_Y, \mathbf{x}_{Cb}, \mathbf{x}_{Cr}\}$ and $\mathbf{y} = \{\mathbf{y}_Y, \mathbf{y}_{Cb}, \mathbf{y}_{Cr}\}$. We propose to use the STSIM for Y component and the original SSIM for Cb and Cr components as follows:

$$\begin{aligned}\text{CSTSIM}(\mathbf{x}, \mathbf{y}) = {} & w_Y \text{STSIM}(\mathbf{x}_Y, \mathbf{y}_Y) + \\ & w_{Cb}\text{SSIM}(\mathbf{x}_{Cb}, \mathbf{y}_{Cb}) + w_{Cr}\text{SSIM}(\mathbf{x}_{Cr}, \mathbf{y}_{Cr}),\end{aligned} \tag{8}$$

where w_Y, w_{Cb}, and w_{Cr} are the weights for each component. In this implementation, we set w_Y=0.6, w_{Cb}=0.2, and w_{Cr}=0.2, respectively. Also, for SSIM, we follow the parameter settings shown in [5]: $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$, $L = 255$, $K_1 = 0.01$, and $K_2 = 0.03$.

4. IMAGE RETRIEVAL SYSTEM IMPLEMENTATION

We have implemented two image retrieval systems for an objective evaluation of the proposed framework; one uses the parsed representation based on the incremental parsing and the other uses the conventional vector quantization for visual information analysis. Both of them are under the same aspect modeling paradigm. The performance of the systems are compared with that of one of the recent image retrieval systems, SIMPLIcity [7], which is based on an image segmentation technique.

4.1. Image database

We implemented the proposed and the benchmark image retrieval systems using 20,000 images obtained from the Corel Stock Photo Library. Also, we identify 15 visual concepts from the database. In many image retrieval systems, each image is considered to fall into only one group. Since each image is a realization of multiple visual sources, we here classify them according to their visual concepts, which are overlapping groups. Among the 20,000 images, there are 9,039 images with one concept, 323 images with two, and 12 with three. All the remaining 10,199 images do not correspond to any

of the visual concepts. Also, we randomly chose 600 query images, each of which contains only one visual concept in order to minimize confusion. We follow the way of the definition of the visual concepts, the number of images, and the number of query images in each visual concept, provided in [3].

4.2. AMPARS system

The proposed image retrieval system parses the given images into a number of variable-size patches by a two-dimensional incremental parsing algorithm. Then, the latent aspects of the occurrence pattern of the parsed representations are trained under PLSA paradigm. The semantic similarity between two images is computed with the likelihood of the images on the concept space formed by the latent aspects.

4.2.1. Two-dimensional Incremental Parsing

Let $\mathbf{X}$ be a two-dimensional vector field taking values from a set of three-dimensional finite vectors. Each element of a vector represents each color component, here red, green, and blue. $\mathbf{X}(\vec{x})$ denotes the symbol vector at the location $\vec{x} \in \mathbb{Z}^2$. Also, we define a subset of $\mathbf{X}$ for an area vector $\vec{a} \in \mathbb{Z}^2$ as follows:

$$\mathbf{X}(\vec{x}; \vec{a}) = \{\mathbf{X}(\bar{x}_1, \bar{x}_2) \,:\, x_i \le \bar{x}_i \le x_i + a_i,\, i = 1, 2\}. \tag{9}$$

Given a dictionary $\mathbb{D}$, we define two operations $|\cdot|$ and $[\cdot]$. $|\mathbb{D}|$ denotes the number of elements of $\mathbb{D}$, $[\mathbb{D}_j]$ refers to an area vector whose element represents the number of pixels of the j^{th} patch along each axis, and $|\mathbb{D}_j|$ corresponds to the number of symbols of the j^{th} patch. At the current pattern matching location, called *anchor point*, denoted by Δ, the set of dictionary indices by ϵ_s-bounded similarity at Δ is

$$\mathbf{H}_s = \{j \mid \mathrm{CSTSIM}_{\max} - \mathrm{CSTSIM}(\mathbb{D}_j,\, \mathbf{X}(\Delta; [\mathbb{D}_j])) \le \epsilon_s, \\ 0 \le j \le |\mathbb{D}|,\, \epsilon \in \mathbb{R}_+\}, \tag{10}$$

where $\mathrm{CSTSIM}_{\max}$ is the maximum value of CSTSIM, equivalent to 1.0, and ϵ_s denotes the bound of the structural similarity. In the proposed implementation, we manually set ϵ_s to 0.015. At each epoch, the parsing scheme constructs the set of indices following the similarity measure. Then, it selects the maximal match index

$$k_{\max} = \underset{k \in \mathbf{H}_s}{\operatorname{argmax}} \{|\mathbb{D}_k|\}. \tag{11}$$

Once the maximal match is found, the scheme appends the dictionary with two new entries, each of which is obtained by appending pixels along the horizontal and the vertical axes. Then, Δ moves following a predetermined heuristic method. In the proposed implementation, a raster scanning order is employed for the movement of Δ.

4.2.2. Aspect Modeling by PLSA

In the aspect modeling of text document, a given text corpus is represented in its empirical distribution $p(w, d)$ by counting the number of words occurred in each document. For an image corpus, we train the aspect model from the empirical distribution of the image-patch observations by PLSA. To generate the empirical distribution, we first have to generate a visual dictionary with which all the images from the corpus can be reconstructed. One feasible heuristic is: at each coding iteration, a given image is encoded with the dictionary updated at the previous iteration. For every n_p image during the encoding procedure, the dictionary entries that are not used for encoding the previous n_p images are pruned, and the dictionary with reduced entries is fed back to the coding step. In the proposed system, we set n_p to 100 and randomly chose 1,200 images for the generation of visual dictionary. The number of entries in the resulting dictionary is 151,390.

Once the visual dictionary is generated, the occurrence patterns of the visual patches in the dictionary are analyzed for the 20,000 images in the database. For learning the latent aspects for a given image corpus by PLSA, the number of latent aspects, K, is manually set to 500. Before learning the PLSA model by EM, $P(z|d)$ and $P(w|z)$ are randomly initialized.

After the latent aspects are modeled, the similarity between the query image and each image in the database is then computed. Generally, the similarity measure between documents under PLSA paradigm is still an open problem. In this implementation, we use the Jensen-Shannon divergence as follows:

$$D_{\mathrm{JS}}(p \,||\, q) = D_{\mathrm{KL}}(p \,||\, m) + D_{\mathrm{KL}}(q \,||\, m), \tag{12}$$

where $m = (p+q)/2$ and D_{KL} is the KL divergence. Also, the similarity between the j^{th} image d_j and the query image d_q is computed as

$$s(d_j,\, d_q) = D_{\mathrm{JS}}(P(z|d_j) \,||\, P(z|d_q)). \tag{13}$$

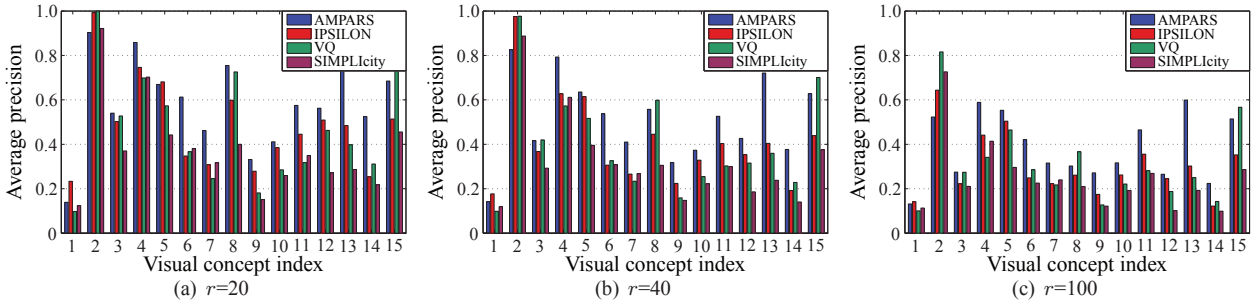
4.3. Fixed-block Representation under PLSA

One considerable difference between the proposed image retrieval system and the conventional systems is the source representation for visual information analysis. As mentioned previously, many of the existing image retrieval systems extract features from fixed-size image blocks or image segments. To compare the performance of the proposed system with that of the conventional approach, we design an image retrieval systems based on the fixed-block image representations trained by VQ as a benchmark system. The benchmark system has the same components as the AMPARS system except one: the visual dictionary is trained by VQ. Each color image is partitioned into 8×8 blocks. Thus, the dimension of the VQ codebook is $8 \times 8 \times 3 = 192$. To train the quantizers, we use the Linde-Buzo-Gray (LBG) algorithm.

5. EXPERIMENTAL RESULTS

We present in this section the retrieval results and the performance evaluation of four image retrieval systems: the proposed AMPARS system, the IPSILON system, the benchmark system, and the SIMPLIcity. IPSILON and AMPARS systems use the parsed representations induced by the incremental parsing algorithms, while the benchmark system is based on the fixed-block representation of visual information. Aspect modeling learned by PLSA technique underlies the benchmark and the AMPARS systems, while the IPSILON systems analyzes the given image corpus by LSA. By the k-means clustering algorithm, the SIMPLIcity partitions a given image into a few regions, then the semantic similarity between the regions of the two given images is computed by an integrated region matching (IRM).

Those systems are implemented with the same image databases and evaluated with the same query images. In many information retrieval, the performance of a retrieval system is commonly evaluated by precision/recall tests. However, in many practical systems, the number of documents (or images) is already innumerable so that the number of retrieved documents is still tremendous. Thus, in this paper, for the evaluation of image retrieval systems, we focus on a few most relevant images without examining the entire retrieved images.



(a) r=20 (b) r=40 (c) r=100

Fig. 1. Comparison of average precisions of the four image retrieval systems for each visual concept.

	AMPARS	IPSILON	VQ	SIMPLIcity
r=20	0.602	0.502	0.420	0.381
r=40	0.550	0.435	0.380	0.323
r=100	0.457	0.353	0.324	0.264

Table 1. Total average precisions of the image retrieval systems.

The retrieval performance is measured by the total average precision. For m visual concepts, the average precision for the r-most relevant precision is $(\sum_{j=1}^{q_i} s_{i,j})/(r \cdot q_i)$, where q_i denotes the number of query images in the i^{th} visual concept and $s_{i,j}$ means the number of relevant images for the j^{th} query in the i^{th} concept. The total average precision for all the queries is defined as $\sum_{i=1}^{m} w_i \sum_{j=1}^{q_i} s_{i,j})/(r \cdot \sum_{i=1}^{m} q_i)$, where $w_i = (m \cdot n_i)/(\sum_{i=1}^{m} n_i)$ and n_i refers to the number of images in the i^{th} visual concept.

Figure 1 compares the average precisions of the four retrieval systems. At r=20, for 11 visual concepts among 15, the average precisions of the proposed AMPARS system are significantly higher than those of the other three systems. Especially, for those concepts, 6, 7, 11, 13, and, 14, the average precisions of the AMPARS system are over 0.1 higher than the other two. Also, at r=40 and r=100, the average precisions of the AMPARS system are considerably higher, particularly for those concepts, 4, 11, and 13. Table 1 provides a numerical comparison of the total average precisions for the four retrieval systems. The total average precisions of the proposed AMPARS system are over 0.10 higher than the other three systems at all the rs, r=20, r=40, and r=100. When compared under the same aspect modeling paradigm, the AMPARS system outperforms the benchmark VQ system in terms of retrieval precision. These comparisons justify that relaxing the constraint of the size of image blocks significantly affects the efficiency of visual information analysis.

6. CONCLUSION AND FUTURE WORK

we have proposed a probabilistic framework of content-based image retrieval based on the parsed representation and implemented the AMPARS system. With a multidimensional incremental parsing technique extended from the Lempel-Ziv incremental parsing, a given image is parsed into a number of patches of variable size. This patch can be thought of as a morphological interface between elementary pixels and a higher level representation than the conventional fixed-block representation. The incremental parsing technique is implemented with a new structural similarity measure that compares the first- and second-order statistics of image patches. By a dictionary generation heuristic approach, a visual dictionary for a given image corpus is generated. For an image corpus, we train the aspect model from the empirical distribution of the image-patch observations by PLSA. The semantic similarity between the given two images is computed by the Jensen-Shannon divergence of two conditional probabilities of the latent aspects. We have implemented two image retrieval systems: one is the proposed AMPARS system, and the other is a benchmark system that uses fixed-block representations of visual information trained VQ. The performance of these two systems is compared with two existing systems: IPSILON based on the incremental parsing with the minimax distortion under LSA paradigm and the SIMPLIcity system based on an image segmentation technique. These four systems are tested with 20,000 images of natural scenes and 600 query images. The experimental results show that the proposed framework captures the visual semantics appeared in the image corpus and outperforms other systems in terms of retrieval precision. This improved performance of the proposed system is due to the two new compoennts, the pattern matching based on the structural similarity and the patch-image analysis by PLSA.

7. REFERENCES

[1] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, no. 5, pp. 530–536, Sept. 1978.

[2] S. H. Bae and B.-H. Juang, "Multidimensional incremental parsing for universal source coding," *IEEE Trans. Image Processing*, vol. 17, no. 10, pp. 1837–1848, Oct. 2008.

[3] S. H. Bae and B.-H. Juang, "Incremental parsing for latent semantic indexing of images," in *Proc. IEEE Int. Conf. Image Processing*, San Diego, CA, Oct. 2008, pp. 925–928.

[4] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, pp. 177–196, Jan. 2001.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visiblity to structural similarity," *IEEE Trans. Image Processing*, vol. 13, no. 8, pp. 600–612, Apr. 2004.

[6] X. Zhao, M. G. Reyes, T. N. Pappas, and D. L. Neuhoff, "Structural texture similarity metrics for retrieval applications," in *Proc. IEEE Int. Conf. Image Processing*, San Diego, CA, Oct. 2008, pp. 1196–1199.

[7] J. Z. Wang, J. Li, and G. Wiederhold, "SIMPLIcity: Semantics-sensitive integrated matching for picture libraries," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 9, pp. 947–963, Sept. 2001.